The Quantification and Visualisation of Human Flourishing

A thesis submitted in partial fulfilment of the requirements for the Degree of

Doctor of Philosophy in

Computational and Applied Mathematical Sciences

by Lisa Henley

Department of Mathematics and Statistics

University of Canterbury

2015

To the two great loves of my life

Craig and Hugh

Table of Contents

1 Acknowledgments

On beginning this journey, I had no idea what a life changing experience it would be. That sounds clichéd, but I really mean it. I feel so fortunate to have been able to do this and humbled by how much my view of the world, and my place within it, has shifted. It is such a privilege to live the line – the more I learn, the less I know.

Additionally, on a physical level my life has changed dramatically in the last four years, most notably as a result of the Canterbury earthquakes. It is difficult to put into words the effect of those events on my life and the life of my family or how poignantly they bought into focus for me "What does it mean - to flourish?"

I am very grateful for the advice and support of my supervisors - Professor Jennifer Brown, Dr David Conradson and Associate Professor Marco Reale. In particular huge thanks to Jennifer whose ability to "cut to the chase" when my work has been stuck or jumbled is nothing short of miraculous. Also for her words of support and encouragement, which have always come at exactly the right time and in exactly the right way – she is an empowerer extraordinaire!

Special thanks to my superstar husband Craig who proof read every word I wrote and provided suggestions, despite knowing that every time it would

result in me crying and shouting – and then (upon reflection) often doing exactly what he suggested. His patience and unfaltering belief in my ability gave me more strength than he could ever realise. Also thanks to our delicious son Hugh who has been so supportive and understanding of the time required throughout this process.

Thanks to my Mum and Dad who have always believed I could do it and who were living examples of an honest day's work, and to my extended family, who have continued to provide a sympathetic ear to my work related ramblings and Facebook postings.

To Kirsty for being the best besty anyone could ever have and for living the highs and lows with me. To my great mates Miriam and Julie, my lovely friend Linda and to my other gorgeous friends, old and new, for their support and encouragement.

Lastly, thanks to Bruce, who sat at my feet consistently for the last two years while I have been working and who has forced me to get out and about to actually do some flourishing on this beautiful Island that we now call home.

2 Abstract

Economic indicators such as GDP have been a main indicator of human progress since the first half of last century. There is concern that continuing to measure our progress and / or wellbeing using measures that encourage consumption on a planet with limited resources, may not be ideal.

Alternative measures of human progress, have a top down approach where the creators decide what the measure will contain.

This work defines a 'bottom up' methodology an example of measuring human progress that doesn't require manual data reduction. The technique allows visual overlay of other 'factors' that users may feel are particularly important.

I designed and wrote a genetic algorithm, which, in conjunction with regression analysis, was used to select the 'most important' variables from a large range of variables loosely associated with the topic. This approach could be applied in many areas where there are a lot of data from which an analyst must choose.

Next I designed and wrote a genetic algorithm to explore the evolution of a spectral clustering solution over time. Additionally, I designed and wrote a genetic algorithm with a multi-faceted fitness function which I used to

select the most appropriate clustering procedure from a range of hierarchical agglomerative methods. Evolving the algorithm over time was not successful in this instance, but the approach holds a lot of promise as an alternative to 'scoring' new data based on an original solution, and as a method for using alternate procedural options to those an analyst might normally select.

The final solution allowed an evolution of the number of clusters with a fixed clustering method and variable selection over time. Profiling with various external data sources gave consistent and interesting interpretations to the clusters.

3 Introduction

3.1 The Evolution Of Progress

The desire to improve is a distinctive and powerful facet of our humanness, from our pursuit to design better tools and to increase our wealth through to our ideas surrounding moral and spiritual progress. Many of the ideas that we associate with the word *progress* were formulated in ancient times (Edelstein, 1967). Throughout the ages however, our ideas about progress have changed. Greek and Roman Philosophers had a large influence on the idea of progress, and later, the development of Christianity added its own contributions, presenting the idea that we could transcend our mortal body to become a heavenly being (Augustine, 1950).

"The education of the human race, represented by the people of God, has advanced, like that of an individual, through certain epochs, or, as it were, ages, so that it might gradually rise from earthly to heavenly things, and from the visible to the invisible."

The idea that progress could be in our own hands, rather than the hands of the gods, emerged later, with Turgot suggesting in the 1700s that there are stages to human progress, each stage superseding the last due to human rather than divine causes (Turgo, 1973). Since that time huge advances in knowledge, developments in science and technology and access to cheap energy have allowed us to grow and 'progress' at an unprecedented rate. We have continued to maintain that progress or growth is 'good' and have largely measured it at an economic level, with Gross Domestic Product (GDP) being the most widely used indicator. Growing the economy is

generally a political election winner and is perceived to be highly correlated, or perhaps even causally linked, to the wellbeing of a nation.

GDP, as we know it today, began life as a tool for guiding the U.S. out of the Great Depression. A 'spreadsheet' combining component parts of the economy, these national accounts were used to assist in production planning during World War II and post war to encourage consumption as a replacement for war production. The tool was not intended to be a measure of a nation's welfare, indeed, Simon Kuznets, one of the early contributors to its development is reputed to have stated, "The welfare of a nation can therefore scarcely be inferred from a measurement of national income…" (U.S. Congress, 1934). GDP misses crucial components of a nations progress and development – "the part that exists outside the realm of monetary exchange" (Rowe, 2008).

Growing the economy of countries where residents are struggling to find food and shelter, generally brings increased life satisfaction to the people (Cummins, 2000). However, there is substantial inequality within and also between countries (Wilkinson & Pickett, 2010), and many people in western societies have more than sufficient food and shelter. Having more than sufficient resources to meet basic needs of food and shelter does not translate directly into satisfaction. Research indicates that once basic needs are met, material wealth contributes to increased life satisfaction at a substantially slower rate (Myers, 2000).

Although this point about the reducing contribution of wealth to life satisfaction is contentious (Hagerty, 2003), there are additional problems with measuring a nation's progress or development using an economic growth based approach. Firstly, there is evidence to suggest that materialistic values actually undermine individual wellbeing (Kashdan, 2007), and additionally, values encouraging continual consumption are in direct opposition to a sustainable future on a planet with finite resources (Matieny, 2000). What will happen when the resources required to fuel the economic growth model, run out?

3.2 Current Ideas On Future Definitions Of Progress

A number of organisations are championing ideas of progress that centre around life satisfaction and planetary wellbeing rather than focusing solely on economic growth. The declaration from the second OECD forum on Statistics, Knowledge and Policy suggested that communities need to consider what progress means in the 21st Century and stated that *"the availability of statistical indicators of economic, social, and environmental outcomes and their dissemination to citizens can contribute to promoting good governance and the improvement of democratic processes. It can strengthen citizens' capacity to influence the goals of the societies they live in through debate and consensus building, and increase the accountability of public policies"* (Istanbul Declaration, 2007). Subsequent to the most recent economic recession, the idea of prosperity without growth was presented in Jackson (2009). The Centre for the Advancement of the Steady State

Economy, founded in 2003, promote *"…the steady state economy as a desirable alternative to economic growth"* (CASSE, 2003). Similarly The Sustainable Europe Research Institute, SERI, (1999) are a *"Pan-European think tank exploring sustainable development options for European societies"*. Richard Layard, Geoff Mulgan and Anthony Seldon founded Action for Happiness in 2010 (AFH, 2010). The organisation's aim is to increase happiness and reduce misery in the world through the way humans approach our lives. They focus on ideas such as developing relationships and community interaction rather than on increasing wealth.

In 2009 the then President of the French Republic, Nikolas Sarkozy, commissioned an investigation to "identify the limits of GDP as an indicator of economic performance and social progress, including the problems with its measurement; to consider what additional information might be required for the production of more relevant indicators of social progress". Subsequently, the Commission on the Measurement of Economic Performance and Social Progress produced a report that defined additional information areas from which to obtain more relevant indicators of social progress (Stiglitz, Sen, & Fitoussi, 2009). The areas identified in this report as being important indicators of social progress are:

1. Material living standards: income, consumption and wealth;

2. Health: mortality, morbidity, mental health

3. Education: a variety of indicators are required to control for inequality

4. Personal activities including work: paid work, unpaid work, commuting, leisure time, homelessness

5. Political voice and governance: participation as full citizen, dissent without fear

6. Social connections and relationships

7. Environment: present and future conditions

8. Insecurity: economic, physical, environmental

The famous phrase from the United States Declaration of Independence (US 1776) "life, liberty and the pursuit of happiness" suggests it is the right of humans to progress towards happiness. The organisations and individuals mentioned here, and others like them, are suggesting alternative ways of creating happier societies that are not dependent solely on that which has long been our primary focus – economic growth.

On April 2 2012 the United Nations (UN) held its first ever Happiness Conference. The report from this conference (Sachs, 2012) states:

The Anthropocene is a newly invented term that combines two Greek roots: "anthropo," for human; and "cene," for new, as in a new geological epoch. The Anthropocene is the new epoch in which humanity, through its technological prowess and population of 7 billion, has

become the major driver of changes of the Earth's physical systems, including the climate, the carbon cycle, the water cycle, the nitrogen cycle, and biodiversity.
The Anthropocene will necessarily reshape our societies. If we continue mindlessly along the current economic trajectory, we risk undermining the Earth's life support systems – food supplies, clean water, and stable climate – necessary for human health and even survival in some places.

The report suggests a new course, with new measures of progress to be developed in line with what they call Sustainable Development Goals. These SDG's are complementary to the Millennium goals, which were developed by the United Nations in 2000 (United Nations, 2000), and aim to ensure poor countries have the right to develop and all countries have the right to happiness. The goals are:

1. End Extreme Poverty
2. Environmental Sustainability
3. Social Inclusion
4. Good Governance

It is important to be clear that the word 'happiness', when mentioned here, is not suggesting a life completely free of negative emotion or experience, but rather life where people are being and doing well, in other words, flourishing. If a country is in a position where increasing wealth is not going to facilitate this goal, then it is indeed time to look at alternatives.

3.3 Current Alternative Measures Of Progress

Below is a description of current developments in the field of alternative measures. It is not exhaustive, but illustrates the range of current approaches.

"Beyond GDP" (an initiative started as a communications platform for the Beyond GDP conference in 2007) are working toward indicators that are complementary to GDP, including the social and environmental aspects of progress. The European Commission, European Parliament, Club of Rome, OECD and WWF hosted their initial conference in 2007 (Beyond GDP, 2007).

Measures are being developed, focusing on varying aspects of being or doing well, where a country can progress toward being or doing well in that area. The Happy Planet Index (New Economics Foundation, 2006), for example, has been produced twice in 2006 and 2009. The index is a single dimension index indicating the ecological efficiency with which a country converts resources to happy life years (HLY). HLY are calculated using subjective life satisfaction scores and life expectancy, and the ecological efficiency indicator is the ecological footprint (EF). Subjective measures of wellbeing are an important dimension of measuring wellbeing, asking people such questions as "All things considered, how well is life going?" gives a good approximation of how well life actually **is** going

(Diener, 1998). The subjective life satisfaction component of the Happy Planet Index is used in a later chapter of this work.

The EF (Global Footprint Network, 2003) was developed in 1990. It measures how much land and water area a population needs to produce the resources it uses and to absorb the carbon dioxide it produces using current technology. At time of writing, using this measure, it takes the earth one and a half years to regenerate what we use in one year.

The Human Development Report was developed in 1990 by Mahbub ul Haq, and The Human Development Index is a major part of this report. The HDI is a composite index consisting in general of Life Expectancy, actual and expected years of schooling and Gross National Income (GNI) per capita. Gross National Income is defined as GDP plus net income received from overseas. The index has been criticised for not including ecological data, however the most current report includes additional, but separate, national data tables and information regarding inequality and environmental sustainability (Human Development Report Office, 2010).

In November 2010, David Cameron, Prime Minister of the United Kingdom asked the Office of National Statistics to develop new measures of wellbeing and progress. Since that time the ONS has been running a public consultation asking questions to discover, "What matters to you?" (The people of the United Kingdom). The themes that emerged from this

consultation were health, relationships, work and the environment. The consultation also highlighted the importance of how people spend their time, and that the importance of a particular theme varies by age and individual. The response suggested the need for a greater sense of fairness and equality. The themes reflected the findings from current research into correlates of wellbeing, but most of the environmental issues raised related to issues such as green space access rather than more global environmental issues. The ONS is currently selecting a number of measures relating to these highlighted themes using such selection criteria as reasonable availability of historical and current data, robustness, suitability for disaggregation and acceptability (as considered by experts in the area). It is anticipated that the list of measures will then be further trimmed using such criteria as stakeholder investment, public acceptance, and sensitivity to public policy intervention. The ONS are examining different approaches to how they will eventually present the information e.g. single index, multiple measures etc. (Office of National Statistics, 2011).

The Better Life Index (OECD, 2011) is an interactive visualisation developed by the OECD. It measures 11 indicators based around two areas - material living conditions and quality of life for all OECD countries. These dimensions were selected based on OECD organisational experience and also research in the area. Users assign weights to each of these indicators based on their subjective view of their importance. This will

allow the OECD to examine what is important to the citizens of each country (based on the dimensions offered). The index doesn't currently include inequality measures within a country.

The Kingdom of Bhutan pursues a development policy of Gross National Happiness (GNH). The emphasis on Happiness as a guide for social policy in Bhutan has historical roots. Social Contracts from 1675 state that "the happiness of all sentient beings, and the teachings of the Buddha are, mutually dependent". The term GNH was coined in the 1970s by His Majesty the Fourth King of Bhutan, Jigme Singye Wangchuck who said, "Gross National Happiness is more important than Gross National Product". Since that time, the wellbeing of the people has been the ultimate development objective in Bhutan (Centre for Bhutan Studies, 2012).

The GNH measure has been developed to reflect the happiness and wellbeing of the people of Bhutan, more accurately than a monetary measure. The measure is used to inform government policy, as well as informing the citizens of Bhutan about current levels of human fulfilment (Centre For Bhutan Studies, n.d.).

The measure is a single number index including nine equally weighted core domains. Each of the domains contains two to four indicators that the Centre for Bhutan studies describe as "statistically reliable, normatively

important, and easily understood by large audiences" (Centre for Bhutan Studies, n.d.) The nine domains are psychological wellbeing, health, time use, education, cultural diversity and resilience, good governance, community vitality, ecological diversity and resilience and living standard. 'Happiness' is defined as having sufficient achievement (termed 'sufficiency' and determined by a cut-off) in 66% of the indicators. Any selection is permitted to allow for diversity. If someone exceeds sufficiency within an indicator, they are capped as having met sufficiency.

The index is calculated as GNH = 1 – HA, where H (headcount) is the proportion of people who do not enjoy sufficiency in 6 or more domains and A is the average proportion of domains in which people who are "not yet happy" lack sufficiency. Examination of results across domains and demographics provides some interesting information. For example from the 2010 Index "In urban areas, 50% of people are happy; in rural areas it is 37%".

3.4 Thesis Outline

In general, current alternative progress measures seem to consist mostly of a single dimension index based on multiple pre-defined indicators that have been identified individually as being important components of the measure. These indicators are usually selected based on scientific research, historical knowledge or particular areas of interest e.g. ecological

concerns. Some measures are asking for user input regarding the importance of each of these indicators, as is the method for the Better Life Index or the UK Wellbeing Index described above. The measures use one, or a small number of indicators as a proxy for the relevant component of a measure. For example carbon footprint, "the total greenhouse gas emissions caused directly and indirectly by a person, organisation, event or product" (The Carbon Trust, 2012), may be used to represent the entire environmental component of a measure.

In this work as a complementary approach, I aim to use statistical techniques, in particular Genetic Algorithms, to define a method for measuring human progress. In contrast to the measures previously discussed, the components (or dimensions) of this measure will not be pre-defined, but rather will be defined from a large number of variables that have been selected for their robustness and close statistical association with one of the areas (outlined earlier) in the Stiglitz-Sen-Fitoussi (2009) report. This report has been a reference point for a number of the progress measures currently being worked on, including the United Kingdom's National Wellbeing work.

I aim to produce a three-dimensional (component) measure to enable ease of graphical visualisation over time (a fourth dimension) for as many countries in the world as there are data available. These dimensions, defined from the statistical methodology, should be both linguistically

interpretable, and should capture the essence of the areas defined in the Stiglitz-Sen-Fitoussi (2009) report.

In this analysis I consider each country as the smallest unit of analysis. A disadvantage of this approach is, that for a particular variable, the information regarding the distribution of the people within that country is lost. For example, Country A may have a higher average life satisfaction value (of 7) than Country B (with a life satisfaction value of 6), but the distribution of the life satisfaction scores of the individuals making up each country's average is unknown. However, this research is attempting to find a global methodology and in datasets appropriate to this end, country is a common unit of analysis. Additionally this research requires data from multiple sources in order to cover the areas found in Stiglitz-Sen-Fitoussi (2009). Using country as the unit of analysis will facilitate the combination of data from multiple sources.

There are many different cultures and cultural practices on earth that may have a positive or negative effect on human progress. This research addresses these differences from the viewpoint that all humans are one species living on one planet, and assumes the following to be true:

> *Some people have better lives than others and these differences relate, in some lawful and not entirely arbitrary way, to states of the human brain and to states of the world. (Harris, 2010)*

I have defined a ‘better life’, in the context of this research, as not only a life that brings more positive feelings (than a ‘worse life’) to the owner of that life, it is a life that is lived in a way that elicits these feelings by meeting needs in a way that does not undermine the planet’s natural systems.

The positive human feelings elicited by a ‘better life’ can be described and measured in a number of ways. It is possible to ask people how happy they are, how content they are or how satisfied they are with their life, and to create measures of life quality or wellbeing. The word ‘happiness’ can have somewhat light, hedonistic connotations in western culture. Indeed it is not hard to find negative response to the ‘Happiness Culture’. The book “Against Happiness: In Praise of Melancholy” is an example (Wilson, 2008).

The Oxford English Dictionary defines ‘well-being’ as the state of being comfortable, healthy or happy (“well-being, n.,” 2015). There are a number of countries currently developing or producing measures of their citizen’s wellbeing. Examples include (as previously mentioned) the United Kingdom, Australia, France and Italy. In New Zealand, Statistics New Zealand has run the General Social Survey since 2008, collecting information on New Zealander’s well-being every two years (Statistics New Zealand, 2009).

For this research, I am using the term human flourishing. This is because, as mentioned, I am attempting to find a method for measuring human progress that not only involves a 'good life', but also living sustainably. This idea of 'human flourishing' embodies the phrase "How do we love all of the children of all of the species for all time" (McDonough, 2005). This represents a state of being that can be measured over time, across all nations. My idea of flourishing is not to necessarily be hedonistically happy all the time, or to even have a high wellbeing if that wellbeing is obtained at the expense of sustainable existence on the planet. Flourishing is defined as human lives lived in a way that maximises life satisfaction and happiness while minimising negative impact on the interconnected earth systems of which we are a part.

GNH is this kind of holistic measure of happiness, however GNH measures happiness as sufficiency in 66% of any of the possible indicators. For example, if you live in the city you may not care about crop loss due to pests so that indicator will not be included in your sufficiency measure. For the purpose of this research the methodology used to define a measure of human flourishing will encompass data from all areas deemed important in Stiglitz-Sen-Fitoussi (2009), irrespective of individual feelings of responsibility. This means that using this methodology, a country may have very high life satisfaction and happiness, but may not be considered flourishing due to low performance on environmental factors.

As an additional stage to this work, I will also examine the relationship between personal values, at a country level, and this new measure of Human Flourishing. This is because personal values and beliefs have a close relationship with our needs, wants and behaviours, which in turn have a complex relationship with our wellbeing (Sagiy & Schwartz, 2000) and can also impact our environment (Banerjee, 1994).

Hadnagy (2011) defines social engineering as *"the art, or better yet science of skilfully manoeuvring human beings to take action in some aspect of their lives"*. Society and its leaders have the power to influence values and actions, and for this reason the aim is to examine what type of values and beliefs correlate with higher levels of human flourishing on various dimensions.

The following sections describe the data used in this work and the methodologies used to reduce the large number of accompanying variables to a subset that is sufficient to account for all the areas of social progress previously discussed. Figure 1 shows this process and may be helpful in understanding the flow of the next few sections.

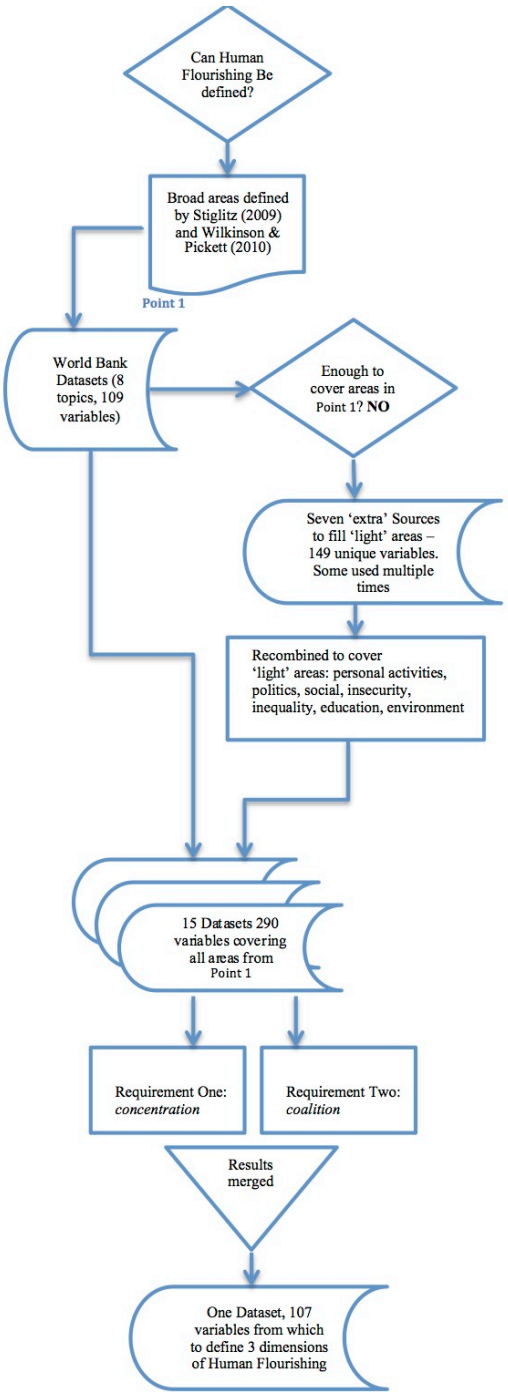


Figure 1: Process flow showing the creation of the flourishing dataset from multiple datasets.

4 The Data

Of central importance in the process of measuring human flourishing are the underlying data and the elements that make up the data. These will largely influence the dimensions of any summary-measure. In order to obtain indicators that would cover all areas outlined, including inequality, I used open source data. I believe this is important in the interest of transparency, reproducibility and accessibility. In order to be included in the work, data needed to be available as a time series for 'all' approximately 200 countries of the world. The exact number of countries in the world differs depending on the source and also changes over time. Each indicator (variable) was examined individually, and if it had more than 25% missing data, it was excluded, as data were considered to not be "missing at random" (Rubin, 1976) and cannot be imputed easily. All datasets were restricted to 1990 onwards as this was the earliest common start point within the data. Of the data sources selected, there is some database infrastructure that is only beginning to be populated.

4.1 Data Selection – Human Flourishing Measure

To reiterate, the following areas are defined as being important in measures of social progress

1. Material living standards

2. Health

3. Education
4. Personal activities including work
5. Political voice and governance
6. Social connections and relationships
7. Environment
8. Insecurity

Source: (Stiglitz et al., 2009)

These areas seem to largely encompass the four sustainable development goals from the UN Happiness Conference mentioned earlier in Sachs (2012). However, one area that I consider missing, or at least underrated is inequality. Inequality has been described as 'The Theory of Everything', and a large volume of evidence for the damaging effects of inequality on social indicators can be found in Wilkinson & Pickett (2010). An end to extreme poverty is one of the sustainable development goals. 'Inequality', although often associated with, is not always related to poverty. It is much more about relative position. For example, a person may have enough money to live on and not be considered poor, yet be unequal to a neighbour who has many times the first person's wealth.

4.1.1 World Bank Data

The World Bank website contains a wealth of data which were opened to the public in 2011 (World Bank, 2011). At that time, the Bank's president Robert Zoellick stated:

"It's important to make the data and knowledge of the World Bank available to everyone. Statistics tell the story of people in developing and emerging countries and can play an important part in helping to overcome poverty" (*World Bank, 2011*).

The data catalogue contains over 2000 indicators for 200 countries from 1960, available by individual indicator, or arranged into topics. The World Bank website states:

"The Primary World Bank collection of development indicators, compiled from officially-recognized international sources. It presents the most current and accurate global development data available, and includes national, regional and global estimates" (World Bank Group, n.d.)

The indicators are compiled from data from many different official sources and there are explanatory notes available on the website for each indicator. The following quote is in response to a user question regarding the way in which the World Bank selects indicators.

"Like many things in life, selecting indicators for the WDI is not an exact science. The intention is to provide good coverage of key development issues, but many of the countries that we work with do not have the quantity - or quality - of data that exists in countries like the United States, for example. Take a look at the Federal Reserve Economic Data (FRED for short); that database alone includes 123,662 economic time-series about the US, including indicators for sub-national areas like states … It's usually not possible to find that level of coverage for indicators in many low and middle income countries. So, while we

follow a set of basic principles, it's relatively rare that we're able to find and publish the perfect indicator: the one that's most relevant for measuring a particular development issue, that's available for every country in the world, for every year, with very high levels of accuracy. More often, indicators have one of more limitations. We try to describe some of these in our metadata and in the sections of the WDI tables called "about the data". You'll find this metadata available in the book, in the on-line tables, and in the databank application - it includes the indicator definition, the source, periodicity, method of aggregation used, statistical concepts and methodology, relevance for development, and limitations and exceptions. The idea behind providing these notes it to help data users decide whether any specific indicator is fit for their purpose. Judgement is required on our part to select an indicator to publish in the WDI; and judgement is required on your part to decide whether it's useful. " (Fantom, 2014)

As such, I accessed the World Bank data by "Topic", selecting each topic for its potential association with the areas outlined in 3.2.

To minimise the limitations described by Fantom (2014) above, the datasets were pre-processed for missing data. The data were firstly restricted to post-1980 in order to match the start point of the World Value Survey data that are used later in this work. The number of missing data points (over the years 1980 to 2010) was calculated for each country and each indicator. After visually assessing the results, each country containing less than 10 missing data points was flagged, for each indicator. Only indicators where the majority of countries were flagged were retained. Following are the topics selected and the number of indicators retained in this way. Descriptions of the 8 datasets used, obtained from the World Bank website, can be found in Appendix 1.

1. Aid effectiveness – 15 indicators,

2. Economic policy and external debt – 24 indicators,

3. Education – 6 indicators,

4. Environment – 19 indicators,

5. Financial sector – 6 indicators,

6. Health – 25 indicators,

7. Labour and social protection – 8 indicators,

8. Urban development – 6 indicators.

The World Bank Data are the main source of data for this work, however during the analysis it was decided that extra sources were needed to ensure full coverage of all areas. The following datasets were selected due to their suitability in meeting this need.

4.1.2 International Labour Organization

The International Labour Organization (ILO) is a specialized agency of the United Nations. It is the only tripartite UN agency bringing "*together representatives of governments, employers and workers to jointly shape policies and programmes promoting Decent Work for all*".

The ILO was founded in 1919 after the First World War with the idea that lasting peace can only be obtained if it is based on Social Justice (International Labour Organization, 1996). The ILO considers that Labour Statistics play an important role in measuring the performance of policies

and progress towards achieving the organisation's "Decent Work for All" agenda. The International Labour Office is the permanent secretariat of the International Labour Organization. The International Labour Office has an open database called LABORSTA containing international labour statistics through to 2008. The database is available online (International Labour Office, 1996a). Metadata for the database is also available online (International Labour Office, 1996b). Indicators in this database are available for individual download by country, topic and publication.

After consideration of the availability of data within data sources yet to be discussed, I downloaded indicators with data available from 1990 to 2008. Consumer Price Index data were not included due to complexity, as they are provided as quarterly information. There are 48 indicators in the database that relate to labour force participation rates. These 48 indicators are participation rates broken down by gender and five-year age bands. In order to prevent this indicator from dominating other areas, this number was reduced to 15 using the method outlined in Fodor (2002). An examination of the remaining indicators showed all had in excess of 40% missing data, and thus were not included.

4.1.3 Indicators Relating To Political Voice And Governance

Professor Bo Rothstein and Professor Sören Holmberg at the University of Gothenburg founded the Quality of Government Institute in 2004. It is a

research institute within the Department of Political Sciences. They conduct research into the causes, consequences and nature of Good Governance and the Quality of Government. One of their aims is to produce cross national, time series data on Quality of Government and its correlates. This is publically available in several formats with accompanying codebook and contains data from 1946 to 2010 (Teorell, 2011). There are 621 variables in the dataset. Once administrative variables and variables with more than 25% missing values were removed there were 55 potentially useful indicators remaining.

Freedom House was founded in 1941 in New York City to encourage support for American involvement in World War II. They state that they are advocates for democracy and human rights around the world. They produce a dataset annually, containing survey information on 195 countries that they have been collecting since 1972. The publically available dataset (Freedom House, 2012) contains survey ratings for political rights, civil liberties and resultant Freedom Status for each country for each year since 1972.

A Polity is most simply understood to be a geographic area and its associated government. The Polity IV project continues the Polity research tradition of analysing the authority characteristics of states in the world for the purposes of analysis. The original Polity Project was undertaken under the direction of Ted Robert Gurr in the 1970s collaborating with Harry

Eckstein. They claim to have: *the most widely used resource for monitoring regime change and studying the effects of regime authority.*

The data are freely available for download and contain information on regime changes and characteristics from 1800 through to 2010. The dataset contains (in addition to other information) the democratic rating (0 – 10), autocratic rating (0-10) and the durability of the government (number of years since there was a regime change).

4.1.4 Further Indicators Of Environmental And Security Concerns

The United Nations Environment Programme produces a publically available data source of 689 indicators, 392 of which contain data available at National level. The data are available online (UNEP, 2006). Each of these indicators needed to be downloaded individually however only 83 contained data within the appropriate time frame. These 83 indicators covered topics regarding climate, disasters, ecosystem management, environmental governance, harmful substances and resource efficiency. There were 23 indicators remaining once those with more than 25% missing data were removed.

The World Health Organisation (WHO) Collaborating Centre for Research on the Epidemiology of Disasters (CRED) has been maintaining an Emergency Events Database since 1988, and this database is available

online for download. It contains information on the location and type of disaster, in addition to the number of people killed and the number of people affected (OFDA/CRED, 1988). The data are at event level, therefore the data were summarised by country and year to obtain the number of people killed and the number of people affected by emergency events.

4.1.5 Further Indicators Of Educational (In)equality

Professor Robert Barro (Professor of Economics, Harvard University) and Jong-Wa Lee (Head of the Asian Development Bank's Office of Regional Economic Integration (OREI) and Acting Chief Economist) developed the Barro-Lee database of Educational Attainment in the World in 1993. The most recent database contains estimates for 146 countries constructed from survey and census data about the distribution of educational attainment broken down by gender and five year age groups (Barro & Lee, J, 2010). There are 48 potentially useful indictors in this dataset.

This results in a dataset of 290 potential indicators, selected from eight sources (the World Bank source containing multiple datasets), from which to form a four dimensional (including time) measure of human flourishing. Details regarding the source and area each of these indicators relates to can be found in Appendix 2.

4.2 Data Selection – Human Flourishing Proxy

The resultant dataset (290 variables) contained a large number of indicators to work with. Section 5 describes the method used to reduce this number to a more manageable level.

This thesis is concerned with human flourishing, defined as a better life, but with the proviso that the life is lived within the constraints of planetary resources. For some of the methods used in the following sections, a dependent variable is required, to represent, or be a proxy for Human Flourishing. As previously described, subjective evaluations of one's life (life satisfaction measures) are recognised as an important measure of wellbeing or happiness. The limitation of using life satisfaction as a proxy for human flourishing is that although it has the potential to capture the "better life" component of human flourishing, it may not capture "within the constraints of planetary resources". This is addressed by individually examining the data relating to each of the areas outlined in 3.2.

There is the potential that there may be areas that have very little relationship with life satisfaction i.e. if you were building a model to predict life satisfaction, these would not necessarily be variables that would be chosen. However, this thesis is not concerned with predicting life satisfaction, it is concerned with selecting the "best of" the variables within all of the areas outlined in 3.2. Therefore, the resultant dataset will include

indicators that are not necessarily closely associated with life satisfaction, but are the most closely associated from their respective areas.

The proxy for human flourishing is only required for one of the two data reduction method described in Section 5, therefore, despite the limitations mentioned, I chose a life satisfaction measure as the proxy. In particular, I chose the subjective life satisfaction index from both the 2006 and 2009 releases of the Happy Planet Index as described in 3.3, which were subsequently matched to the relevant years of the data under investigation.

For the 2006 release of the Happy Planet Index, the life satisfaction information came from a wide number of sources, and values for countries where no data were available were estimated by the HPI developers using modelling techniques. The 2006 HPI is available for 179 countries. In the 2009 release, the life satisfaction information was obtained from responses to the satisfaction with life questions in the Gallup World Poll and World Values Survey. The 2009 HPI is available for 143 countries. The index documentation states that statistical modelling techniques were applied to take into account differences between the two surveys to ensure that the life satisfaction data used to build the final index were comparable to the first, but does not mention what those techniques are (Abdallah, Michaelson, Shah, Stoll, & Marks, 2012).

I initially examined the possibility of using another subjective index as the proxy - positive and negative effect, which is a composite of people's evaluations of their feelings. It would have been useful to have taken this route and to have examined potential negative drivers of flourishing, however a lack of data made this impractical.

4.3 Data Selection – Personal Values

For the second part of this work, examining the association between personal values and the newly defined human flourishing measure from the first part of this work, the World Values Survey will be used. The World Values Survey in collaboration with the European Values Study is a survey that has been carried out worldwide five times since 1981. A nationally representative sample of respondents are interviewed in each of 97 societies (90% of the world's population) using a standardised questionnaire and asked to share information regarding their values and beliefs *"concerning religion, gender roles, work motivations, democracy, good governance, social capital, political participation, tolerance of other groups, environmental protection and subjective well-being"* (World Values Survey, 2008). The 1081 variable dataset is freely available online for download and this dataset will be used to examine the association between personal values and human flourishing.

4.4 Summary

A number of open data sources were utilised to obtain indicators associated with areas deemed important in defining future measures of social progress. World Bank data were the primary data source, producing 109 indicators from 8 topics or datasets. Seven further data sources were required in order to cover all 'important areas'. These provided a further 181 indicators, producing 290 indicators in all.

A human flourishing proxy was required for some of the statistical methods used. For this purpose the life satisfaction indicator from the Happy Planet Index was chosen.

Finally, the World Values Survey data will be used to examine the association between the human flourishing measure and personal values.

5 Reducing A Dataset

The requirements to have a subset of variables that best summarise and describe the full dataset are common to most exercises of feature selection. Additionally, the need to conserve data storage space and processing time encourages selective data inclusion (Guyon, 2003).

The 290 variables were arranged into datasets to align with the areas outlined in 3.2. At this point there were more variables than there were countries available for analysis within each year, i.e. the variable to observation ratio exceeded 1. Therefore, I chose to undertake a data reduction exercise with the following requirements.

Requirement 1: Find the ideal subset of variables which best summarised each dataset. This approach will be referred to as "concentration" and does not require a proxy for human flourishing.

Requirement 2. Find the ideal subset of variables from each dataset that is most closely associated to the proxy for human flourishing (life satisfaction). This approach will be referred to as "coalition".

Requirement 1 is necessary as requirement 2 necessitates a circular process - attempting to find the variables most closely associated with the proxy for flourishing, however human flourishing has not yet been measured. As mentioned, life satisfaction cannot be considered a complete proxy of

human flourishing. Therefore, a flexible methodology was defined with a separate technique to examine each of the requirements, a process comparable to the recommendation in Guyon (2003). The following section addresses requirement 1.

5.1 Concentrating A Dataset: A More Traditional Approach

5.1.1 Introduction

Linear Principal Component analysis (PCA) involves taking a matrix (X_{nm}) of m measurements on n individuals or objects and creating p new variables or principal components where $p < m$. This is done in such a way that the first principal component explains as much variability in the data as possible; the second explains the next largest amount and so on. Ideally, a large proportion of the information in the data can be explained by a small number of the new variables or principal components, thereby reducing the data.

If the data are not distributed normally a transformation such as taking the log or ranking the data, is sometimes performed. Rank transformation is particularly useful for data with outliers and unusual distributions (Baxter, 1995).

PCA is usually undertaken to create a reduced number of new dimensions that summarise the data. An alternative method for reducing the number of

variables or dimensions in a dataset is referred to by Fodor (2002). This method involves calculating the principal components of a dataset, then using the scree plot to decide on the number of original variables to keep (as opposed to principal components). A scree plot is a line plot showing the eigenvalue λ_i, of each of the p principal components on the x-axis, and the component number, i, on the y-axis. It is possible to calculate the proportion of variance in the data that each of the p components explains using equation (1).

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \quad \textbf{(1)}$$

The principal components are ordered by decreasing eigenvalue, so a large difference in two adjoining points on the plot, relative to other adjoining points, shows as a ‘break’ or ‘kick’. The number of variables to retain using this method is decided in a similar way to deciding the number of Principal Components (PC) to be retained. It may be decided upon examination of the Scree Plot, Figure 3, that those PCs with eigenvalues less than one should be discarded, or to discard PCs that come after a 'break' or ‘kick’ as described.

Then, beginning with the smallest eigenvalue, the variable that has the largest absolute coefficient in the corresponding eigenvector is removed. The process is repeated (without replacement) until the required number of variables remains as determined by the scree plot examination.

5.1.2 Method

As mentioned, the main data source used in the development of the human flourishing measure is the World Bank Data described in Carroll (1972). Beginning with the Health Dataset described in Section 6, the distributions of the data were examined. The data were not normally distributed. There were outliers, some variables appeared to be distributed with a somewhat exponential distribution and others not (see Figure 2). To help with these issues and for consistency, the data were rank transformed, that is, replaced by the value of their ranks. In this case, the data contain ties. Ranks were averaged within ties, as shown in Table 1. Observations with missing values were treated as passive observations, and they were not included in the subsequent analysis. The scale of the variables being used affects the results of Principal Components Analysis. If there are no ties in the data, ranking the data means all variables have the same scale (Baxter, 1995). However, because there are ties and missing observations in the data, the correlation matrix was calculated and used to calculate the principal components.

Obs	Original	Rank
1	1,000	5
2	1,000	5
3	1,000	5
4	1,000	5
5	1,000	5
6	1,000	5
7	1,000	5
8	1,000	5
9	1,000	5
10	10,000	10.5
11	10,000	10.5
12	100,000	12

Table 1 An example of averaging within ties

The Fodor (2002) implementation of PCA described in 5.1 was then applied firstly to the World Bank Health Dataset described in 4.1.1 to develop the macro and test the methodology, and then to each remaining World Bank Dataset, to select the number of required variables to sufficiently summarise each dataset. The World Bank Health Dataset was chosen for methodological testing, as this is the second item listed in 3.2.

Once this process was complete, the variables that had been selected from each dataset were combined into one dataset and it was determined to which of the human progress areas (4.1) each one belonged.

It was apparent that some areas, for example ‘personal activities’ and ‘education’ had fewer variables than others, so at this point I looked for other data sources to ensure relatively equal number of variables associated

with each area. I looked for data that were specifically related to the lacking areas, for example the Barro-Lee database of Educational Attainment in the World (BLD). All of these extra data are described in 4.1.2 though 4.1.5.

Some of the datasets had variables or indicators potentially common to multiple areas of social progress. For example the BLD contains variables stratified by gender and age and is therefore useful for potential indicators of inequality. For this reason the variables from the extra datasets were recombined from their original datasets into new datasets related to each of the lacking areas. This meant that a particular variable could be part of multiple datasets. This allows that once the reduction process is performed a given variable may not 'make it through' in one area but may do so in another.

The data were transformed as described, and then the Principal Components from each dataset were used as described in 5.1.1 to decide on the variables required to sufficiently summarise each dataset.

5.1.3 Results

Figure 2 shows the substantial variety in the frequency distributions of the individual variables within the World Bank Health Dataset (used for the methodological setup), supporting the decision to transform the data.

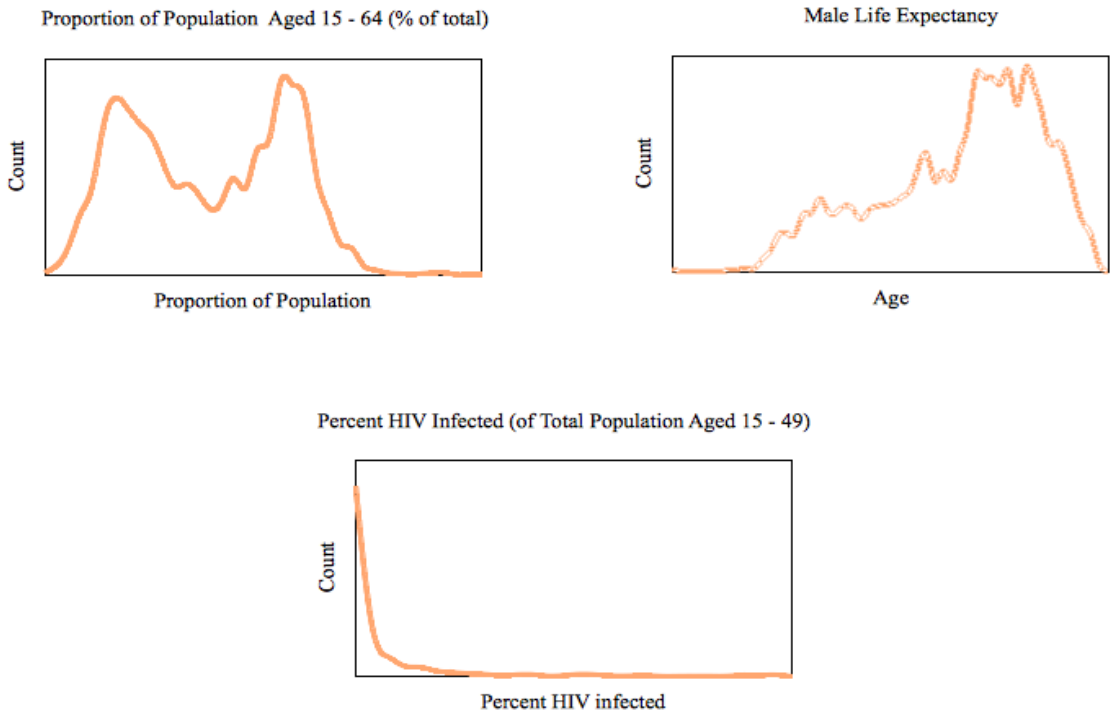


Figure 2: A sample of frequency distributions from variables within the World Bank health dataset showing the wide variety of distribution shape.

Figure 3 shows the scree plot from the PCA being applied to the rank-transformed World Bank Health dataset. The plot suggests it is appropriate to retain 5 variables. Although the principal components are not being used directly, as an aside the first 5 components in the model explain 81% of the variance within the data. Scree plots of this sort were produced for all of the 15 datasets under investigation (8 World Bank and 7 'lacking areas').

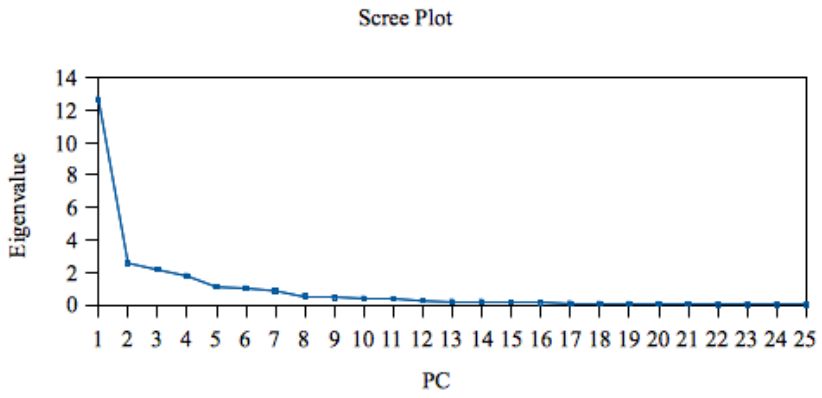


Figure 3: The Scree plot from the PCA of the rank-transformed World Bank health dataset.

For each of the 15 datasets a table was produced showing the order in which each variable should be discarded according to the result from the Fodor (2002) implementation of PCA. An example using the World Bank Health dataset is shown in Table 2. There are instances where a variable that appears early on (due to a strong association with an unimportant

principal component), can also be found later on to have a strong association with a relatively important principal component. Fodor does not discuss how to handle these variables, however the methodology was followed as written in Fodor (2002), as there were few instances where this occurred and this is not the only method to be used as selection criteria. The method to find the *coalition* will be described in the following section. Variables below the solid line in Table 2 are the set of retained variables. Variable names within the datasets are not necessarily easily inferred and may contain prefixes or suffixes not directly related to the individual variables.

Drop Order First to Last	
Variable Name	**Description**
SP_DYN_CBRT_IN	Birth rate, crude (per 1,000 people)
SP_DYN_LE00_MA_IN	Life expectancy at birth, male (years)
SP_POP_0014_TO_ZS	Population ages 0-14 (% of total)
SH_XPD_PUBL_ZS	Health expend, public (% of GDP)
SH_XPD_TOTL_ZS	Health expend, total (% of GDP)
SH_XPD_OOPC_TO_ZS	Out-of-pocket health expend (% of total expend on health)
SP_DYN_CBRT_IN	Birth rate, crude (per 1,000 people)
SP_POP_65UP_TO_ZS	Population ages 65 and above (% of total)
SH_XPD_PCAP_PP_KD	Health expend per capita, PPP (constant 2005 international $)
SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)
SH_TBS_INCD	Incidence of tuberculosis (per 100,000 people)
SH_XPD_PUBL_GX_ZS	Health expend, public (% of government expend)
SH_XPD_EXTR_ZS	External resources for health (% of total expend on health)
SP_POP_GROW	Population growth (annual %)
SH_TBS_INCD	Incidence of tuberculosis (per 100,000 people)
SP_POP_TOTL_FE_ZS	Population, female (% of total)
SH_TBS_DTEC_ZS	Tuberculosis case detection rate (%, all forms)
SH_DYN_AIDS_ZS	Prevalence of HIV, total (% of population ages 15-49)
SH_XPD_OOPC_ZS	Out-of-pocket health expend (% of private expend on health)
SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)
SP_POP_TOTL	Population, total
SP_DYN_CDRT_IN	Death rate, crude (per 1,000 people)
SH_XPD_PRIV_ZS	Health expend, private (% of GDP)
SH_XPD_OOPC_TO_ZS	Out-of-pocket health expend (% of private expend on health)
SP_POP_0014_TO_ZS	Population ages 0-14 (% of total)

Table 2 A sample table showing the order in which variables were discarded using the PC's and procedure described by Fodor.

Variables retained include private expenditure on health, the out of pocket proportion of private expenditure on health, death rates and total population.

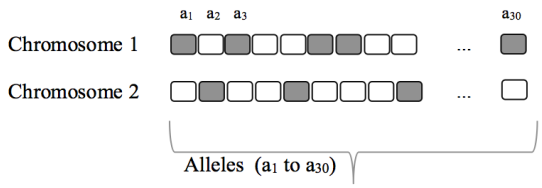
5.1.4 Summary

To satisfy the areas defined as being important in defining measures of Social Progress, 15 datasets were compiled containing a total of 290 variables. The Principal Components of each dataset were used as described in 5.1.1 to satisfy the requirement to "find the ideal subset of variables which best summarised each dataset". The traditional implementation of PCA forms new dimensions or variables that summarise the dataset, this implementation chooses the 'best' variables, in their current state, to summarise the dataset. This methodology has a limitation in that a variable may be associated with both important and unimportant principal components, however this occurred infrequently and this method is one of two methods being used for feature selection. The comparative technique is described in the next chapter, using a Genetic Algorithm to find the *coalition* - the ideal subset of variables from each dataset that is most closely associated with human flourishing. Full details of the variables chosen by method one for each of the 15 datasets are not shown here as the variables that were ultimately chosen by comparing and contrasting the two approaches can be seen in Table 10.

5.2 Genetic Algorithms – A Brief Introduction

Genetic Algorithms are iterative, heuristic (experience based) search processes that can be used to find solutions to problems where an exhaustive search of all potential solutions would be impractical due to time or resource constraints. John Holland (1993) was credited with their invention in the early 1970s. They mimic natural evolution by using techniques such as natural selection, inheritance, mutation and crossover. They are used in fields and contexts where the optimal solution is required from within a large search space. In this research I am attempting to find the best subset of variables to summarise a larger dataset.

Typically, a Genetic Algorithm (GA) starts with an initial population that represents the potential solution space. This initial population consists of chromosomes that are traditionally a string of zeros and ones, and is usually a randomly generated sample of the solution search space. For example, for a variable selection / reduction exercise where there are 30 variables to choose from, each chromosome will be 30 bits long. Each of these bits is called an allele and each of these 30 alleles will consist of either a randomly generated 0 (indicating the variable is not to be selected) or 1 (indicating the variable is to be selected). Figure 4 is a visual example of two chromosomes, each 30 alleles long, from a population encoded to represent variable selection or non-selection from a total of 30 variables.



Variable (Allele) Name	Chromosome 1	Chromosome 2
Variable 1	Included	Not Included
Variable 2	Not Included	Included
Variable 3	Included	Not Included
Variable 4	Not Included	Not Included
Variable 5	Not Included	Included
Variable 6	Included	Not Included
Variable 7	Included	Not Included
Variable 8	Not Included	Not Included
Variable 9	Not Included	Included
…	…	…
Variable 30	Included	Not Included

Figure 4 Example chromosomes each representing 30 alleles indicating whether one of 30 variables should be selected or not

The size of the initial population of chromosomes is only one of the control parameters that must be decided for a genetic algorithm. Recommendations in this area can suggest that the size of the initial population should represent the size of the problem space. There are suggested methods for finding optimal control parameters, for example Grefenstette (1986) suggests a two-stage process selecting firstly an appropriate GA to solve the problem and then secondly using further algorithms to find the optimal control parameters. Alternatively, as in Vinterbo (1999), a variety of parameters can be tried and the results evaluated in context. For example, increasing the population size may result in a better solution, but will also result in increased processing time.

The next step is to choose an appropriate fitness function to evaluate each of the chromosomes. The fitness function is important as it determines how the 'best' or maximum of the solution space is measured. For example, if the aims of the analysis were to find the best subset of potential independent variables in terms of explaining a dependent variable, as well as creating the simplest subset possible, then Akaike's Information Criterion (AIC) may be a suitable measure as AIC can meet both of these aims (Bozdogan, 1987). If however, simplicity was not paramount, and there was more interest in the explanatory power of the variable subset, then Root Mean Square Error (Beal, 2005) may be the appropriate fitness function.

The fitness function is evaluated for each member of the initial population. Using the example given, with 30 potential independent variables, and choosing the AIC as the fitness function, we would build a regression model for each member of the initial population. We would use the alleles with a value of 1 within each chromosome to decide which independent variables are in the model. Once the model has been run, the resultant AIC would be stored with the chromosome and this becomes the 'fitness' of that chromosome or solution.

Once the members of the initial population have had their fitness calculated, a selection method is chosen to decide which individual will carry on to the next step in the process. This can be done using a number

of methods including, for example, 'roulette-wheel selection', where there is a chance for all to be selected, but a greater chance of selection for individuals of greater fitness, or 'tournament selection' where a number of individuals compete in a tournament, or competition, and the winner, the fittest individual, is selected to proceed to the next step. The total number of individuals chosen to move onto the next stage is another control parameter in the algorithm.

The winning chromosomes then undergo crossover and mutation. This is the method by which new potential solutions are generated. There are a number of crossover methods, the simplest is to randomly choose two parents from the winning group and swap their alleles from a random point on their chromosome, to create two new offspring. The offspring may then undergo mutation that can also happen in a number of ways. The simplest involves reversing the value of a random bit in a very small number of chromosomes, for example, changing a 0 to a 1 for 0.25% of chromosomes. Mutation helps to avoid local maxima (chromosomes with good, but not optimal fitness) and premature termination of the process, but the rate of mutation should be low. Figure 5 is a pictorial representation of the process of crossover and mutation. Two chromosomes shown one on top of the other swap a section of alleles (A and B) at the crossover point. Section A, (which was formerly attached to

the upper chromosome) is now attached to the bottom chromosome and the final allele of this section is subsequently mutated.

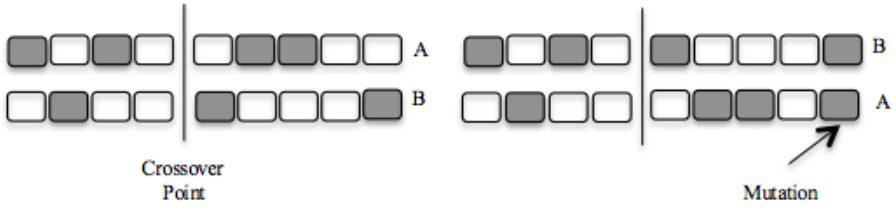


Figure 5: A pictorial representation of the crossover and mutation processes within the Genetic Algorithm.

The newly formed offspring are then evaluated in terms of their fitness, combined with the parent population, and a new selection process occurs. This selection, crossover, mutation, evaluation loop continues until the 'best' solution is reached. The point of termination can be decided in a number of ways, for example, if the average fitness of the offspring does not change for 20 generations and the solution could therefore be judged 'stable'.

In the previous section of this work an application of Principal Component Analysis was used to find the *concentration* - the ideal subset of variables that best summarised each dataset.

In this next section a GA is used in conjunction with Regression Analysis in order to find the *coalition*.

5.3 Finding The Coalition Within A Dataset

5.3.1 Introduction

Regression is a tool for analysing the relationship between a dependent variable, and one or more independent variables. It is often used for predictive modelling, but can also be used to determine which of the independent variables are most closely associated with the dependent variable. There are multiple forms of regression analysis, and the form chosen can have a substantial impact on the result. For example, using stepwise selection in a regression analysis may produce different results to an analysis on the same data using forward selection (Zhang & Xu, 2001). Therefore, even with classical regression, there is a large potential solution space to be explored (2^n), where n is the number of variables to choose from. In this work I designed and wrote a GA, to sample the solution space in order to find an optimal solution. The data under investigation were the World Bank and 'extra data' sources, the same data used in the *concentration* approach, described in 4.1.2 to 4.1.5.

Use of regression requires a dependent variable, in contrast to PCA, and for this role I selected the Human Flourishing proxy, life satisfaction, described in 4.2. As life satisfaction is not expected to fully explain human flourishing, I will consider the results of both the *concentration* and the *coalition* methods, and additionally, will analyse the data associated with each of the areas in 3.3 individually.

The Genetic Algorithm attempts to find the subset of variables that, together, have the closest relationship with the dependent variable. This subset will then become the 'best' summary of the dataset.

5.3.2 Method

A residual analysis was carried out on a test dataset to check the suitability of using linear regression.

Beginning with the World Bank Health dataset and then proceeding to the remaining World Bank and extra data sources described in 4.1.2 through to 4.1.5, the human flourishing proxy (life satisfaction) was matched by country and year to the relevant records of each dataset. Each of the 15 datasets were analysed individually. The human flourishing proxy, life satisfaction, is available for two years only, however, as mentioned, the aim of this part of the analysis was to reduce the original number of variables from which to build the human flourishing measure and visualisation, of which life satisfaction is but a part.

I wrote an algorithm, containing parameters to allow multiple re-runs for each of the 15 datasets. The parameters were initial chromosome population size, chromosome length (number of independent variables), number of replicates being sampled from the population, number of replicates selected for crossover (the fittest x), number of elites (chromosomes retained from one generation to another without crossover),

a ‘stable’ value (the number of iterations the average fitness function must remain the same before stability is deemed to be reached), maximum iterations (to prevent continuous running if no stable solution could be found), selection method (forward, backward or stepwise), significance level of (variable) entry and significance level of (variable to) stay. Although there are further, alternate forms of regression, for the sake of implementation simplicity, only the three mentioned were included in the analysis.

Firstly an initial randomly generated population of chromosomes was created, each chromosome consisting of 0/1 alleles. Each chromosome was of a length equal to the number of independent variables in the dataset to be reduced. Population size of either 100 or 200 was used here to check consistency of results.

The population was randomly sampled without replacement in its entirety to create a number of replicates equal to half of the population. Then a regression equation, equation (2), was built for each member of each replicate (binary tournament), using the variables selected by the alleles of the chromosome. Only main effects were included.

$$E[Y] = b_0 + b_1X_1 + b_2X_2 + \cdots + b_p X_p, where\ p \leq q \qquad \textbf{(2)}$$

where *Y* is the response, or dependent, variable, and the proxy for human flourishing, life satisfaction; the *X*s represent the *p* explanatory variables selected by the Genetic Algorithm using forward, backwards or stepwise regression; the *b*s are the regression coefficients, and *q* is the number of variables selected by the initial chromosome. The Root Mean Square Error (RMSE) was selected as the fitness function, as the aim was to find the subset of variables (within each dataset) that had the most explanatory power in regards to the dependent variable. This fitness function is another parameter that can be easily changed in the code. The algorithm checks which of the two tournament opponents has the lowest RMSE (fitness function), and retains that chromosome as the fittest (winner of the tournament). It performed this operation for all of the pairs to obtain the 'best' of each replicate.

The fittest *x* replicates were selected for crossover as per the previously set parameter. A copy was taken of this group, as they became the ‘parents’ of the next generation. Then the average fitness of the parent group was retained in a macro variable. I chose a ‘stable’ parameter of 20, meaning the average fitness of the group needed to remain stable for 20 iterations before termination of the algorithm.

At this point, (for each iteration), the two fittest chromosomes were preserved as ‘elites’. Then all of the remaining tournament winners, or fittest were re-sampled without replacement, again into replicates of

sample size two. Crossover was performed at a random point along the chromosomes (the same point for each member of a pair and in 0.25% of cases mutation took place.

A regression model was again built for each member of each tournament pair and the fitness calculated. The fittest from each pair was retained and this group was combined with the parents of that generation. Again, the fittest (parameterised) x of the group was selected for crossover and the process continued for either 500 generations or until the average fitness was stable.

As an example, shown in Figure 6, a population of 100 would divide into 50 replicates of sample size two, regression models would be built for each member of each group, and the fitness calculated. Forty (crossover parameter) of the fittest chromosomes would be selected for crossover. These would be copied (they are the parents), the average fitness of the whole group calculated and stored and the two fittest preserved as elites. The remaining 38 chromosomes would be re-sampled into 19 groups of two. A regression model built for each, the fitness calculated and the fittest 20 recombined with the 40 parents. The top 40 would be selected for crossover and the process begins again (copying the parents) until stability is reached.

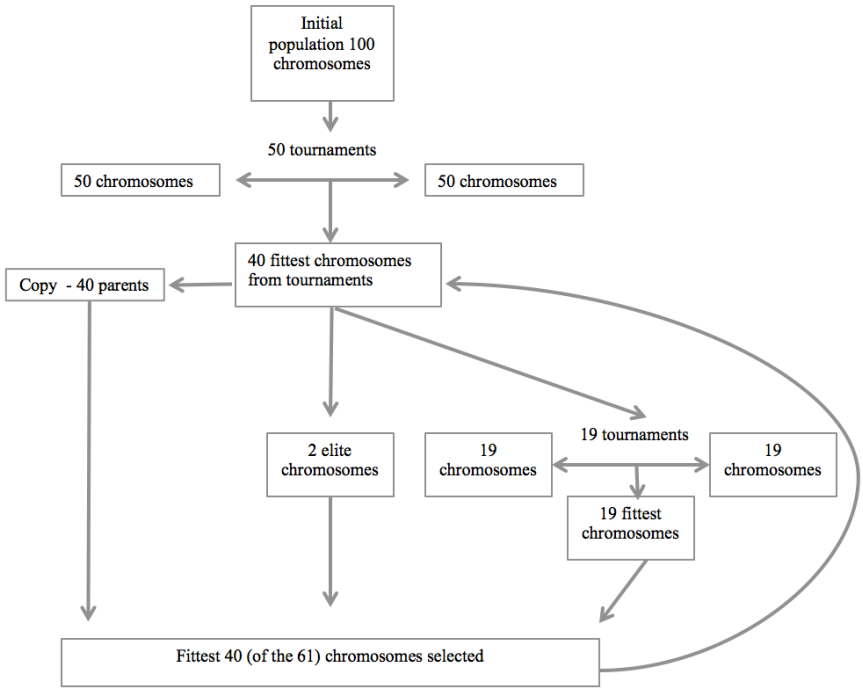


Figure 6 An example GA population evolution

The algorithm was run multiple times for each of the 15 datasets and varying population size, crossover amount and the regression type. The significance level for entry (SLE) of a variable into the model was set to 0.20 and the significance level for a variable to stay in the model (SLS) was set to 0.15. The aim of this analysis was to choose variables to be subsequently analysed, it is not a predictive analysis, and I did not wish to exclude potentially valuable variables.

5.3.3 Results

Tables similar to that shown in Table 3 were produced for each of the 15 datasets.

Trial	Population	Crossover Count	Elite Count	Time (minutes)	Generations	R-Squared	Variables Kept	Variable Description
Forward Selection								
1	100	40	2	3	55	0.707167	SH_IMM_IDPT	Immunization, DPT (% of children ages 12-23 months)
							SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)
							SH_XPD_EXTR_ZS	External resources for health (% of total expenditure on health)
							SH_XPD_OOPC_TO_ZS	Out-of-pocket health expenditure (% of total expenditure on health)
							SH_XPD_PCAP_PP_KD	Health expenditure per capita, PPP (constant 2005 international $)
							SH_XPD_PUBL_GX_ZS	Health expenditure, public (% of government expenditure)
							SP_DYN_CBRT_IN	Birth rate, crude (per 1,000 people)
							SP_DYN_LE00_FE_IN	Life expectancy at birth, female (years)
							SP_DYN_LE00_MA_IN	Life expectancy at birth, male (years)
							SP_POP_0014_TO_ZS	Population ages 0-14 (% of total)
							SP_POP_65UP_TO_ZS	Population ages 65 and above (% of total)
2	100	40	2	3	55	0.707167	As above	
3	200	80	2	6	58	0.70559	As above	
4	200	80	2	7	58	0.70559	As above	
Stepwise selection								
5	100	40	2	10	188	0.70655	SH_IMM_IDPT	Immunization, DPT (% of children ages 12-23 months)
							SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)
							SH_XPD_EXTR_ZS	External resources for health (% of total expenditure on health)
							SH_XPD_OOPC_TO_ZS	Out-of-pocket health expenditure (% of total expenditure on health)
							SH_XPD_PCAP_PP_KD	Health expenditure per capita, PPP (constant 2005 international $)
							SH_XPD_PUBL_GX_ZS	Health expenditure, public (% of government expenditure)
							SP_DYN_CBRT_IN	Birth rate, crude (per 1,000 people)
							SP_DYN_LE00_FE_IN	Life expectancy at birth, female (years)
							SP_POP_0014_TO_ZS	Population ages 0-14 (% of total)
							SP_POP_65UP_TO_ZS	Population ages 65 and above (% of total)
6	100	40	2	10	188	0.70655	As above	
7	200	80	2	18	160	0.702371	As Above	
8	200	80	2	15	160	0.702371	As above	

Backward selection								
9	100	40	2	5	92	0.704446	SH_IMM_IDPT	Immunization, DPT (% of children ages 12-23 months)
							SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)
							SH_XPD_EXTR_ZS	External resources for health (% of total expenditure on health)
							SH_XPD_OOPC_TO_ZS	Out-of-pocket health expenditure (% of total expenditure on health)
							SH_XPD_PCAP_PP_KD	Health expenditure per capita, PPP (constant 2005 international $)
							SH_XPD_PRIV_ZS	Health expenditure, private (% of GDP)
							SH_XPD_PUBL	Out-of-pocket health expenditure (% of private expenditure on health)
							SH_XPD_PUBL_GX_ZS	Health expenditure, public (% of government expenditure)
							SH_XPD_TOTL_ZS	Health expenditure, total (% of GDP)
							SP_DYN_CBRT_IN	Birth rate, crude (per 1,000 people)
							SP_DYN_LE00_FE_IN	Life expectancy at birth, female (years)
							SP_POP_0014_TO_ZS	Population ages 0-14 (% of total)
							SP_POP_65UP_TO_ZS	Population ages 65 and above (% of total)
10	100	40	2	5	92	0.704446	As above	
11	200	80	2	13	98	0.702777	As above	

Table 3: A sample summary table, in this case for the World Bank health dataset, of details regarding algorithm runs and resultant variables retained.

In the example shown in Table 3, using the World Bank Health Dataset, forward and stepwise regression produced similar results, selecting variables related to health expenditure, immunisation rates, life expectancy, birth rates and the proportion of young and elderly people within the population. The only difference between the forward and stepwise techniques was that male life expectancy was eliminated in the latter. The backward elimination technique chose the same variables, but added additional expenditure variables to the mix.

The R-squared values differed for each dataset. For example the R-squared of the World Bank Health analysis, shown in Table 3, were relatively good compared to the 'extra data' environment dataset, shown in Table 3. The variables in that dataset had almost no explanatory power for the human flourishing proxy.

This was to be expected and was the reason why each dataset was analysed separately. As mentioned, life satisfaction was only a proxy for human flourishing and it was important that variables from all of the areas outlined in 3.2 be included in the final dataset.

Trial	Population	Crossover Count	Elite Count	Time (minutes)	Generations	R-Squared	RMSE	Variables Kept	Variable Description
Forward Selection									
1	100	40	2	13	101	0.37	0.9195	NY_ADJ_DKAP_GN_ZS	Adjusted savings: consumption of fixed capital (% of GNI)
								NY_ADJ_AEDU_GN_ZS	Adjusted savings: education expenditure (% of GNI)
								NY_ADJ_DCO2_GN_ZS	Adjusted savings: carbon dioxide damage (% of GNI)
								NY_GDP_NGAS_RT_ZS	Natural gas rents (% of GDP)
								NY_ADJ_AEDU_CD	Adjusted savings: education expenditure (current US$)
								NY_ADJ_DFOR_GN_ZS	Adjusted savings: net forest depletion (% of GNI)
2	100	40	2	13	101	0.37	0.9195	As above	
3	200	80	2	38	143	0.35	0.9182	NY_ADJ_DKAP_GN_ZS	Adjusted savings: consumption of fixed capital (% of GNI)
								NY_ADJ_AEDU_GN_ZS	Adjusted savings: education expenditure (% of GNI)
								NY_ADJ_DCO2_GN_ZS	Adjusted savings: carbon dioxide damage (% of GNI)
								NY_ADJ_DCO2_CD	Adjusted savings: carbon dioxide damage (current US$)
								NY_GDP_TOTL_RT_ZS	Total natural resources rents (% of GDP)
4	200	80	2	39	143	0.35	0.9182	As above	
Stepwise selection									
5	100	40	2	9	64	0.18	0.9213	NY_ADJ_DCO2_GN_ZS	Adjusted savings: carbon dioxide damage (% of GNI)
								NY_GDP_NGAS_RT_ZS	Natural gas rents (% of GDP)
								NY_ADJ_AEDU_CD	Adjusted savings: education expenditure (current US$)
6	100	40	2	9	64	0.18	0.9214	As above	
7	200	80	2	67	218	0.3	0.9174	NY_ADJ_DKAP_GN_ZS	Adjusted savings: consumption of fixed capital (% of GNI)
								NY_ADJ_AEDU_GN_ZS	Adjusted savings: education expenditure (% of GNI)
								NY_ADJ_AEDU_CD	Adjusted savings: education expenditure (current US$)
Backward selection									
9	100	40	2	16	98	0.25	0.9176	NY_ADJ_AEDU_GN_ZS	Adjusted savings: education expenditure (% of GNI)
								NY_ADJ_DCO2_GN_ZS	Adjusted savings: carbon dioxide damage (% of GNI)
								NY_ADJ_DMIN_CD	Adjusted savings: mineral depletion (current US$)
10	100	40	2	16	98	0.25	0.9176	As above	

Table 4 Genetic Algorithm runs for the World Bank environment dataset

It is important to note that the variables selected by each chromosome are not necessarily those that end up as 'selected' in the table. The chromosome defines the 'starting' dataset, the forward, backward or stepwise process then determines what remains at the end.

The full results of all 15 datasets are shown in Table 10 when the results from the *concentration* analysis are compared with the results from the *coalition* analysis.

5.3.4 Summary

I designed and wrote a Genetic Algorithm to determine the 'best' subset of variables for each of the 15 datasets, in terms of their association with the human flourishing proxy. The GA used the RMSE from regression models to determine the 'fitness' of each tested subset of variables. Changing the type of regression, or other macro parameters, can produce varying results (although they are similar in the example case of the World Bank Health dataset). Additionally, the datasets have different levels of explanatory power in terms of the human flourishing proxy, life satisfaction. It would be interesting to examine a time-lag regression of the 'extra data' environmental data onto the human flourishing proxy to examine if environmental changes have an impact on human flourishing when they have had a longer chance to take effect.

In the next implementation of a GA I will create a fitness function with more than one component.

The next section details the comparison of the PCA results from both the *concentration* and the *coalition* analyses and the resultant complete dataset from which the dimensions of human flourishing will be determined.

5.4 The Comparison

In 5.1 an application of Principal Component Analysis was used to find the *concentration* - the ideal subset of variables that best summarises each dataset. In the last section 5.3, a GA was used in conjunction with Regression Analysis in order to find the *coalition*.

When the results of both the PCA application and Genetic Algorithm approaches are combined for each dataset, there is a range of overlap in the variable selection, where each variable can be 'selected by none' through to 'selected by all'. Table 5 shows a sample from the World Bank dataset comparisons. As an example of the range of selection options, pre primary school enrolment (SE_PRE_ENRR) in the education dataset is selected very strongly by 10 runs of the GA but is not selected at all by the PCA. It is for this kind of reason that I chose to use a "majority rules" approach to selection. A variable was kept if, either the PCA method selected it, or if more than two runs of the GA selected it. Table 6 shows the number of World Bank variables selected in this way. The WB dataset column shows

the World Bank dataset they were selected from (Table 7) and the area of social development column indicates the area of social development they are related to as per Table 8.

			Genetic Algorithm		
PCA	**Variable Name**	**Variable Label**	**Variable Name**	**Variable Label**	**Runs**
Aid Effect					
✔	DT_ODA_ODAT_GI_ZS	Net offocial development assistance received (% of gross capital formation)	DT_NFL_UNCR_CD	Net official flows from UN agencies, UNHCR (current US$)	9
✔	DT_ODA_ALLD_KD	Net official development assistance and official aid received (constant 2008 US$)	DT_ODA_ODAT_MP_ZS	Net ODA received (% of imports of goods and services)	8
✔	EN_ATM_CO2E_PC	CO2 emissions (metric tons per capita)	EN_ATM_CO2E_PC	CO2 emissions (metric tons per capita)	3
✔	IT_CEL_SETS_P2	Mobile cellular subscriptions (per 100 people)	IT_CEL_SETS_P2	Mobile cellular subscriptions (per 100 people)	6
Economic Policy					
✔	NE_DAB_TOTL_CN	Gross national expenditure (current LCU)	NE_IMP_GNFS_ZS	Imports of goods and services (% of GDP)	2
✔	NE_EXP_GNFS_ZS	Exports of goods and services (% of GDP)	NE_RSB_GNFS_ZS	External balance on goods and services (% of GDP)	3
✔	NE_EXP_GNFS_ZS	Exports of goods and services (% of GDP)	NY_ADJ_DKAP_GN_ZS	Adjusted savings: consumption of fixed capital (% of GNI)	7
✔	NE_IMP_GNFS_ZS	Imports of goods and services (% of GDP)	NY_ADJ_DMIN_CD	Adjusted savings: mineral depletion (current US$)	7
✔	NE_RSB_GNFS_ZS	External balance on goods and services (% of GDP)	NY_ADJ_DNGY_GN_ZS	Adjusted savings: energy depletion (% of GNI)	7
✔	NY_ADJ_DNGY_GN_ZS	Adjusted savings: energy depletion (% of GNI)	NY_GDP_DEFL_ZS	GDP deflator (base year varies by country)	5
	NY_GDP_MKTP_CN	GDP (current LCU)	NY_GDP_MKTP_KD	GDP (constant 2000 US$)	2
✔	NY_GDP_MKTP_PP_KD	GDP, PPP (constant 2005 international $)	NY_GDP_MKTP_KD_ZG	GDP growth (annual %)	3
✔	NY_GDP_PCAP_CD	GDP per capita (current US$)	NY_GDP_MKTP_PP_KD	GDP, PPP (constant 2005 international $)	2
✔	NY_GDP_PCAP_KD	GDP per capita (constant 2000 US$)	NY_GDP_PCAP_CD	GDP per capita (current US$)	3
			NY_GDP_PCAP_KD	GDP per capita (constant 2000 US$)	7
Education					
✔	SE_ENR_PRIM_FM_ZS	Ratio of female to male primary enrollment (%)	SE_ENR_PRIM_FM_ZS	Ratio of female to male primary enrollment (%)	2
✔	SE_PRM_AGES	Primary school starting age (years)	SE_PRE_ENRR	School enrollment, preprimary (% gross)	10
✔	SE_PRM_ENRL	Primary education, pupils	SE_PRM_AGES	Primary school starting age (years)	6
			SE_PRM_DURS	Primary education, duration (years)	8

Table 5: A sample variable list showing the multi-method - that is both PCA and Genetic Algorithm - approach to variable selection.

Area Soc Dev	Variable Name	Variable Description	WB Data set
1	DT_ODA_ODAT_GI_ZS	Net official development assistance received (% of gross capital formation)	2
1	DT_ODA_ALLD_KD	Net official development assistance and official aid received (constant 2008 US$)	2
7	EN_ATM_CO2E_PC	CO2 emissions (metric tons per capita)	2
1	IT_CEL_SETS_P2	Mobile cellular subscriptions (per 100 people)	2
1	NE_IMP_GNFS_ZS	Imports of goods and services (% of GDP)	3
1	NE_RSB_GNFS_ZS	External balance on goods and services (% of GDP)	3
1	NY_ADJ_DKAP_GN_ZS	Adjusted savings: consumption of fixed capital (% of GNI)	3
7	NY_ADJ_DMIN_CD	Adjusted savings: mineral depletion (current US$)	3
1	NY_GDP_MKTP_PP_KD	GDP, PPP (constant 2005 international $)	3
1	NY_GDP_PCAP_CD	GDP per capita (current US$)	3
1	NY_GDP_PCAP_KD	GDP per capita (constant 2000 US$)	3
3	SE_ENR_PRIM_FM_ZS*	Ratio of female to male primary enrolment (%)	4
3	SE_PRE_ENRR	School enrolment, pre-primary (% gross)	4
3	SE_PRM_AGES	Primary school starting age (years)	4
3	SE_PRM_DURS	Primary education, duration (years)	4
3	SE_PRM_ENRL	Primary education, pupils	4
3	NY_ADJ_AEDU_CD	Adjusted savings: education expenditure (current US$)	6
3	NY_ADJ_AEDU_GN_ZS	Adjusted savings: education expenditure (% of GNI)	6
7	NY_ADJ_DCO2_GN_ZS	Adjusted savings: carbon dioxide damage (% of GNI)	6
7	NY_ADJ_DNGY_GN_ZS	Adjusted savings: energy depletion (% of GNI)	6
7	NY_GDP_TOTL_RT_ZS	Total natural resources rents (% of GDP)	6
1	BX_KLT_DINV_CD_WD	Foreign direct investment, net inflows (BoP, current US$)	7
1	FM_AST_DOMS_CN	Net domestic credit (current LCU)	7
4	SL_TLF_CACT_ZS	Labour participation rate, total (% of total population ages 15+)	10
4	SL_TLF_TOTL_IN	Labour force, total	10
7	SP_URB_TOTL_IN_ZS	Urban population (% of total)	16
2	SH_IMM_IDPT	Immunization, DPT (% of children ages 12-23 months)	8
2	SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)	8
2	SH_XPD_EXTR_ZS	External resources for health (% of total expenditure on health)	8
2	SH_XPD_OOPC_TO_ZS	Out-of-pocket health expenditure (% of total expenditure on health)	8
2	SH_XPD_PCAP_PP_KD	Health expenditure per capita, PPP (constant 2005 international $)	8
2	SH_XPD_PRIV_ZS	Health expenditure, private (% of GDP)	8
2	SH_XPD_PUBL_GX_ZS	Health expenditure, public (% of government expenditure)	8
2	SP_DYN_CBRT_IN	Birth rate, crude (per 1,000 people)	8
2	SP_DYN_CDRT_IN	Death rate, crude (per 1,000 people)	8
2	SP_DYN_LE00_FE_IN	Life expectancy at birth, female (years)	8
2	SP_DYN_LE00_MA_IN	Life expectancy at birth, male (years)	8
2	SP_POP_0014_TO_ZS	Population ages 0-14 (% of total)Ê	8
2	SP_POP_65UP_TO_ZS	Population ages 65 and above (% of total)	8
7	SP_POP_TOTL	Population, total	8
9	SL_TLF_CACT_MA_ZS	Labour participation rate, male (% of male population ages 15+)	15
9	SL_TLF_CACT_FE_ZS	Labour participation rate, female (% of female population ages 15+)	15
9	SL_TLF_TOTL_FE_ZS	Labour participation rate, female (% of total population ages 15+)	10

Table 6: Variables selected from the complete World Bank dataset by PCA method, PCA method and GA, or by the GA

Description	WB Dset
Aid Effect	2
Economic Policy	3
Education	4
Environment	6
Financial Sector	7
Health	8
Labour	10
Urban Development	16

Table 7: The areas of World Bank data examined. Each area is held as a separate dataset.

Area of Social Development	**# Vars**
1. Material living standards	11
2. Health	13
3. Education;	7
4. Personal activities including work	2
5. Political voice and governance	0
6. Social connections and relationships	0
7. Environment (present and future conditions	7
8. Insecurity, of an economic as well as a physical nature	0
9. Inequality	3

Table 8: The areas of social development being examined.

From the examination of Table 8 the decision was made to obtain extra data in some of the areas of social development as some areas had fewer variables than others at this point. These areas were education, personal activities, political voice and governance, social connections and relationship, environment, insecurity: economic as well as physical and inequality.

A sample of the 'extra data' runs is shown in Table 9, indicating some variables are selected by one technique and not the other (for example FPR_55N59 – Female Labour Force Participation rate for 55 – 59 year olds), by both (MPR_20N24) or relatively strongly by the GA (FPR_65P).

1. Personal Activities		
Variable Name	**PCA**	**GA**
FPR_25N29		2
FPR_55N59	1	
FPR_65P		10
MFPR_25N29		4
MFPR_30N34		6
MFPR_35N39	1	
MFPR_40N44		2
MFPR_45N49		2
MFPR_50N54		2
MFPR_55N59	1	6
MPR_20N24	1	10
MPR_50N54		8
MPR_50N54		2
MPR_65P		6

Table 9: A sample of 'extra data' variable selection results showing both genetic algorithm runs and selection by PCA.

When all analysis runs were complete, tables such as those shown in Table 9 were examined and the relevant variables were selected based on their selection by the PCA technique, the PCA and multiple GA runs, or selection by more than two GA runs. All of these resultant variables were combined into one dataset shown in Table 10. The variables were re-examined and then associated with the one or more areas of social development (Table 8). This means that a variable can potentially fill multiple roles. The totals for each of these areas are shown at the bottom of the Table. There are some areas that have more members than others, but there are enough in each area to proceed.

5.5 Summary

One hundred and seven variables were selected from 290 variables obtained from World Bank and other international datasets to provide information regarding important areas of social development (Table 8).

There is a great deal of work associated with compiling a dataset of this sort. In particular, a given source of data may initially look very promising, but further investigation will identify large gaps in the data, particularly for historical data. It would be easier to attempt to build a dataset using data from just one point in recent history rather than a time series, but this removes the opportunity to examine changes over time. Data are not always stored in the most access-efficient manner, for example some of the environmental data were available one variable at a time and it was only once a variable was downloaded that its use could be determined.

There were two requirements for the reduction of the data. Firstly an application of PCA was used to 'summarise' each dataset and secondly I designed and wrote a Genetic Algorithm to find the variables within a dataset that were most relevant to human flourishing.

The application of PCA was simple to use but, as described previously, there were issues when a variable was correlated with both an important and an unimportant Principal Component. The advantage of using a

Genetic Algorithm to find an optimal solution is that the user is largely in control of how 'success' is defined. This GA was written primarily for the selection of variables most closely associated with the human flourishing proxy. It searched a much larger solution space than an approach where an analyst might choose to use one of stepwise, forward and backward regression and a manual selection of variables. Although this work used a combination of two techniques (the PCA method and the GA) to select variables, it was an exploratory kind of analysis, and it would be easy to convert the GA to a predictive analysis searching a much larger solution space than normal for the 'best' solution. In this way a Genetic Algorithm is very flexible.

Variable	Description	1	2	3	4	5	6	7	8	9
_eco__agri_area_	Agricultural Area							1		
_gen__exchange_r	Local Currency Units per $US	1						1		
_gen__fertility_	Number of Children per woman	1					1	1		
_gen__infant_mor	Infant Deaths per 1000 births	1	1					1	1	
_gen__land_area_	Land Area							1		
_gen__life_expec	Life expectancy	1	1					1		
_gen__migrants_n	Number of Migrants 000s	1				1	1	1	1	1
_gen__migration_	Number of Migrants per 1000 people	1				1	1	1	1	1
_gen__mobile_pho	Mobile Subscriptions per 100 inhabitants	1					1	1		
_gen__pop_rural_	Rural Population as percent of total	1			1			1		
_gen__pop_total_	Total Population 000s							1		
_gen__pop_urban_	Urban population as percent of total	1			1			1		
_res__agri_prod_	Agricultural Production Index Base 1999-2001 - Total							1		
_res__cereals_ha	Cereals Area Harvested kms							1		
_res__cereals_yi	Cereals Yield Hectograms per Hectare							1		
_res__fish_catch	Fish Catch Metric tons							1		
affected	Total affected by a given disaster	1					1	1	1	
al_religion	Religious fractionalisation QOGOV						1			1
BX_KLT_DINV_CD_WD	Foreign direct investment, net inflows (BoP, current US$)	1								
chga_hinst	Regime Institutions 0 democ 6 royal dictator					1		1	1	
dpi_cemo	Chief Executive a Military Officer					1		1	1	
dpi_checks	Number of Veto Players QOGOV					1		1	1	
dpi_lipc	Legislative Index of Political Competitiveness QOGOV					1		1	1	
dpi_numul	Number of Seats non-aligned/allegiance unknown QOGOV					1		1	1	
dpi_system	RegimeType QOGOV					1		1	1	
DT_ODA_ALLD_KD	Net official development assistance and official aid received (constant 2008 US$)	1								
DT_ODA_ODAT_GI_ZS	Net official development assistance received (% of gross capital formation)	1								
durable	Number of years in power POLITY					1			1	
EN_ATM_CO2E_PC	CO2 emissions (metric tons per capita)							1		
F15_Prim_Comp	Percent Female Pop 15+ Complete Primary			1	1					1
F15_Prim_Tot	Percent Female Pop 15+Went to Primary			1	1					1
F15_Sec_Tot	Percent Female Pop 15+ Went to Secondary			1	1					1

F15_Year_Tert_School	Female 15+ Avg Years Tert Schooling			1	1					1
F25_No_Schooling	Percent Female Pop 25+ No Schooling			1	1					1
F25_Sec_Tot	Percent Female Pop 25+ Went to Secondary			1	1					1
F25_Year_Prim_School	Female 25+ Avg Years Prim School			1	1					1
F25_Year_Tot_School	Female 15+ Avg Years Tot Schooling			1	1					1
fe_etfra	Ethnic fractionalisation QOGOV						1		1	1
fe_plural	Plurality Group QOGOV						1			
fh_cl	civil liberties QOGOV					1			1	
fh_pr	Political rights QOGOV					1			1	
FM_AST_DOMS_CN);	Net domestic credit (current LCU)	1						1		
FPR_25N29	Female Labour Force Participation Rate 25 to 29				1				1	1
FPR_55N59	Female Labour Force Participation Rate 55 to 59				1				1	1
FPR_65P	Female Labour Force Participation Rate 65 +				1				1	1
gd_ptss	Political Terror Scale – US State Department QOGOV					1			1	
ht_colonial	Colonial Origin QOGOV					1				
ht_region	The Region of the Country QOGOV					1				
ht_regtype1	Simplified Regime Type QOGOV					1				
IT_CEL_SETS_P2	Mobile cellular subscriptions (per 100 people)	1						1	1	
killed	People killed in given disaster	1						1	1	1
lp_catho80	Religion Catholic as % pop 1980 QOGOV				1		1			
lp_lat_abst	Latitude						1			
lp_legor	Legal Origin						1			
lp_muslim80	Religion Muslim as % pop 1980 QOGOV				1		1			
lp_no_cpm80	Religion Other Denom as % pop 1980 QOGOV				1		1			
lp_protmg80	Religion Protestant as % pop 1980 QOGOV				1		1			
MF15_Pop_N_000s	Total Pop 15+	1		1				1		1
MF15_Prim_Comp	Percent Total Pop 15+ Complete Primary			1						1
MF15_Prim_Tot	Percent Total Pop 15+ Went to Primary			1						1
MF15_Sec_Tot	Percent Total Pop 15+ Went to Secondary			1						1
MF15_Year_Tert_School	Total Pop 15+ Avg Years Tert School			1						1
MF25_No_Schooling	Percent Total Pop 25+ No Schooling			1						1
MF25_Year_Prim_School	Total Pop 25+ Avg Years Prim School			1						1
MFPR_25N29	Total Labour Force Participation Rate 25 to 29				1				1	1
MFPR_30N34	Total Labour Force Participation Rate 30 to 34				1				1	1
MFPR_35N39	Total Labour Force Participation Rate 35 to 39				1				1	1

MFPR_40N44	Total Labour Force Participation Rate 40 to 45				1				1	1
MFPR_45N49	Total Labour Force Participation Rate 45 to 49				1				1	1
MFPR_50N54	Total Labour Force Participation Rate 50 to 54				1				1	1
MFPR_55N59	Total Labour Force Participation Rate 55 to 59				1				1	1
MPR_20N24	Male Labour Force Participation Rate 20 to 24				1				1	1
MPR_30N34	Male Labour Force Participation Rate30 to 34				1				1	1
MPR_50N54	Male Labour Force Participation Rate 50 to 54				1				1	1
MPR_65P	Male Labour Force Participation Rate 65+				1				1	1
NE_IMP_GNFS_ZS	Imports of goods and services (% of GDP)	1								
NE_RSB_GNFS_ZS	External balance on goods and services (% of GDP)	1								
NY_ADJ_AEDU_CD	Adjusted savings: education expenditure (current US$)			1				1		
NY_ADJ_AEDU_GN_ZS	Adjusted savings: education expenditure (% of GNI)			1				1		
NY_ADJ_DCO2_GN_ZS	Adjusted savings: carbon dioxide damage (% of GNI)	1						1		
NY_ADJ_DKAP_GN_ZS	Adjusted savings: consumption of fixed capital (% of GNI)	1								
NY_ADJ_DNGY_GN_ZS	Adjusted savings: energy depletion (% of GNI)	1						1		
NY_GDP_MKTP_PP_KD	GDP, PPP (constant 2005 international $)	1								
NY_GDP_PCAP_CD	GDP per capita (current US$)	1								
NY_GDP_PCAP_KD	GDP per capita (constant 2000 US$)	1								
NY_GDP_TOTL_RT_ZS	Total natural resources rents (% of GDP)	1						1		
SE_ENR_PRIM_FM_ZS	Ratio of female to male primary enrolment (%)			1						1
SE_PRE_ENRR	School enrolment, pre-primary (% gross)			1						
SE_PRM_AGES	Primary school starting age (years)			1						
SE_PRM_DURS	Primary education, duration (years)			1						
SH_IMM_IDPT	Immunization, DPT (% of children ages 12-23 months)	1	1							
SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)	1	1							
SH_XPD_EXTR_ZS	External resources for health (% of total expenditure on health)	1	1							
SH_XPD_OOPC_TO_ZS	Out-of-pocket health expenditure (% of total expenditure on health)	1	1							
SH_XPD_PCAP_PP_KD	Health expenditure per capita, PPP (constant 2005 international $)	1	1							
SH_XPD_PRIV_ZS	Health expenditure, private (% of GDP)	1	1							
SH_XPD_PUBL_GX_ZS	Health expenditure, public (% of government expenditure)	1	1							
SL_TLF_CACT_MA_ZS	Labour participation rate, male (% of male population ages 15+)				1					1
SL_TLF_CACT_ZS	Labour participation rate, total (% of total population ages 15+)				1					1
SL_TLF_TOTL_FE_ZS	Labour participation rate, female (% of female population ages 15+)				1					1
SL_TLF_TOTL_IN	Labour force, total				1					1
SP_DYN_CBRT_IN	Birth rate, crude (per 1,000 people)	1	1					1		

SP_DYN_CDRT_IN	Death rate, crude (per 1,000 people)	1	1					1		
SP_DYN_LE00_FE_IN	Life expectancy at birth, female (years)	1	1						1	
SP_DYN_LE00_MA_IN	Life expectancy at birth, male (years)	1	1						1	
SP_POP_0014_TO_ZS	Population ages 0-14 (% of total)Ê		1					1		
SP_POP_65UP_TO_ZS	Population ages 65 and above (% of total)		1					1		
SP_POP_TOTL	Population, total							1		
SP_URB_TOTL_IN_ZS);	Urban population (% of total)	1						1		
wdi_fr	Fertility Rate births per woman QOGOV		1	1			1			
		39	18	25	36	20	21	46	41	48

Table 10: The complete flourishing dataset variable list in addition to the area of social development each variable is associated with.

6 Defining the Flourishing Landscape

Creating dimensions that will cross culture, gender, age and all other demographic groupings, to constitute a life well lived seems at first glance like an arduous task. In fact, it could be that Human Flourishing is impossible to define in a way that would satisfy everyone. This task could be seen as similar to that presented by Sam Harris with reference to morality in his book The Moral Landscape (2010). He maintains that science can determine 'what is moral?' similar to the way 'what is healthy?' can be scientifically determined. The definition of a healthy life is not the same for everyone and it is continually developing as new information becomes available, but we would be unlikely to find someone who thinks that having a serious illness is better than not having a serious illness. This chapter proceeds with the corollary that the same applies to Human Flourishing – that we can scientifically create a current and dynamic definition of that which constitutes A Good Life.

The previous sections describe the way in which 107 variables were assembled into a dataset. These variables represent the 'best' indicators, relating to the areas outlined in 3.2, from which to define a measure of human progress or flourishing. In the following sections I will describe the application of several statistical techniques to summarise and enable visualisation of the dataset.

6.1 Principal Components Analysis

6.1.1 Introduction

In the previous section 5.1, an application of Principal Components Analysis was used to choose the individual variables which 'best' summarised a larger dataset, as described by Fodor (2002). In this section, a more traditional application of PCA is implemented.

6.1.2 Method

The principal components of the 107 variable condensed dataset were calculated, using data from the year 1995. Only 1995 data were used, subsequent years will be dealt with in later chapters. The year 1995 was chosen because, as shown in Table 11, more than half of the variables within this year had no missing data.

Missing %	Number of variables
0	62
1	18
2	1
4	8
6	1
7	6
8	2
10	2
11	2
16	1
26	1
28	1
31	1
34	1

Table 11 Number of variables with the associated percentage of missing data.

This dataset will be referred to as CD_1995. Before calculating the principal components missing data were imputed using the median value. Although this is not desirable, it is preferred over the alternative of leaving in missing values. With missing values, listwise deletion is used in this case, and with this, a record with even one missing data point is deleted. Listwise deletions would result in only $1/5^{th}$ of the data being available for processing. The data, with imputed missing values, were replaced by the value of their ranks for consistency with the previous implementation of Principal Components Analysis.

The scree plot was used to decide the appropriate number of principal components required to explain a reasonable portion the variation in the data as described in 5.1.1. The retained PCs were then examined to determine if they could be explained by the variables used to create them.

6.1.3 Results

The scree plot Figure 7, indicated a break at six principal components. This is a large number of dimensions in terms of visualisation of the data. However, as the first six principal components (six with the largest eigenvalues) explained 65% of the variation in the data, six were extracted for examination.

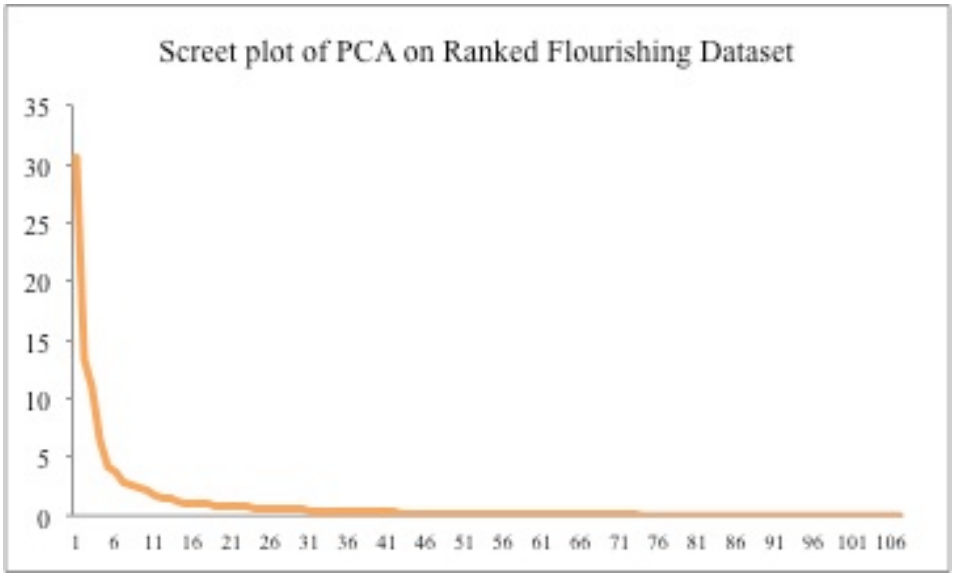


Figure 7: The scree plot from a PCA on the complete ranked CD_1995 dataset.

The twenty variables with the highest absolute correlation with each principal component were examined. This number was chosen as beyond this point the correlations became relatively weaker. In this way the six components can be interpreted as indicated in Table 12.

Component	Description
1: Healthcare, Education and CO2 emissions	Positively correlated with years tertiary schooling (in particular women), mobile phones, total female years of schooling 25+, GDP per capita, health expenditure, CO2 emissions, life expectancy male and female Negatively correlated with proportion of children in population, birth-rate, elderly working, proportion with no schooling, external funds for healthcare as % of total healthcare spend
2: Population and Land size	Positively correlated with population in rural areas, population in general, arable land, crops harvested, GDP PPP, people killed in natural disasters, labour force Negatively correlated with number of women, Imports of goods and services (as a % of GDP)
3: Labour force participation	Positively correlated with labour force participation for 15-49 year olds, particularly women Negatively correlated with religious diversity
4: Political Freedom	Negatively correlated with political competitiveness Positively correlated with Muslims as % of population, lack of rights and freedom
5: Primary Education	Positively correlated with primary education
6: Primary education for women	Positively correlated with primary education, in particular for women

Table 12: A description of the first six components from the PCA on the ranked CD_1995 dataset.

A plot of the first two principal components for each country is shown in Figure 8. Moving across the plot from left to right would tend to bring increased average amount of tertiary schooling (in particular for women),

more mobile phones, higher GDP per capita, higher life expectancy for both males and females, but also increased CO2 emissions. In the same direction across the plot, birth rate and proportion of children in the population would tend to decrease, there are less elderly people working, less people with no schooling and less external help with healthcare spending. The countries that are to the right of the vertical axis include economically developed nations such as Germany Sweden, The Netherlands, Norway, New Zealand and the United States of America. Wealthy Middle Eastern countries such as Kuwait and Bahrain also appear to the right of this axis.

Moving up the plot increases the contrast between general population numbers and female population numbers, as the population increases the relative proportion of women decreases. Imports of goods and services, as a proportion of GDP tends to decrease moving up the plot. Also moving up the plot, arable land, crops harvested, GDP PPP, the labour force and people killed in natural disasters all tend to increase. Countries that have a relatively large land size appear at the top of this axis, such as the United States of America, India, Russia and China. Small island nations such as Fiji, Tonga and the Maldives appear at the bottom of this axis.

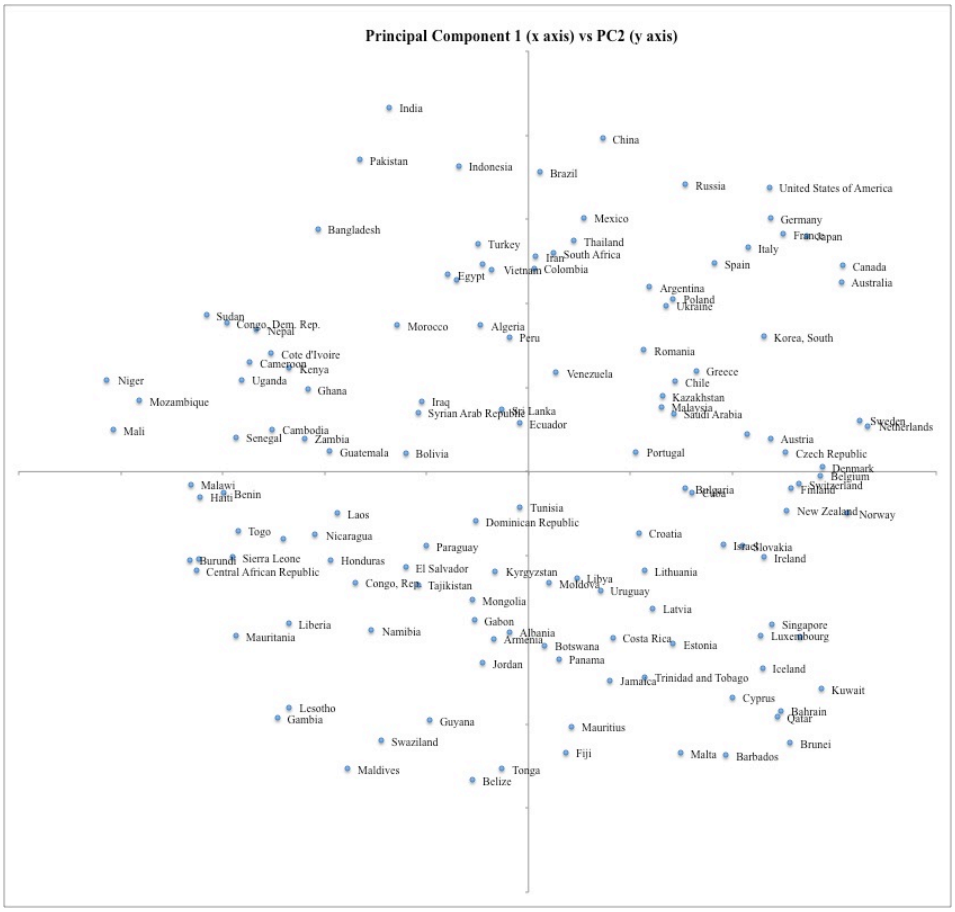


Figure 8: A plot of the first two components, for each country, from the PCA on the ranked CD_1995 dataset.

6.1.4 Summary

Figure 8 is a demonstration of the way that principal components are often visualised. The two principal components displayed here can be interpreted and they explain 41% of the variation in the data. The two principal components are not explaining a large amount, however there are a large number of variables in the data set (107 in total). The first two dimensions appear to focus on contrasts in economic development and land size.

Adding further dimensions to this type of plot, while explaining more of the variation in the data, will also add further complexity to the visualisation. Possible improvements to visualisation will be explored in later chapters.

6.2 Cluster Analysis

An alternative way of summarising the dataset and subsequently enabling visualisation of the dataset is to find groups of countries that share similar traits within a particular group or cluster, and vary significantly between clusters (Jain, 1999). Each cluster can then be profiled or interpreted by those traits and potentially visualised over time. This allows an implementation of the landscape view of human flourishing mentioned in the introduction to the chapter 6, where there will be potential peaks and valleys with respect to a better or worse life. The following sections describe the application of three clustering approaches. The first is an application of Spectral Clustering. Next a Genetic Algorithm (GA) is designed, written and used to select the 'best' clustering method from a range of different methods. These two approaches will then be compared to the K-Means procedure.

6.2.1 Spectral Clustering

Humans examining an image or scene can easily segment it into areas or groups. Spectral clustering arose relatively recently in the field of

computer vision from the need to do this grouping automatically (Weiss, 1999). The spectral clustering approach has a number of advantages over traditional clustering techniques. It often produces better results with regards to cluster detection and is technically simple to undertake (Von Luxburg, 2007).

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with vertex set $V = \{v_1 \dots v_n\}$. Each edge between two vertices v_i and v_j has an associated non-negative weight $w_{ij} > 0$. The *weighted adjacency matrix* of $\mathcal{G}$ is $W = [w_{ij}]$, where $i, j = 1, \dots, n$. Two vertices v_i and v_j are connected if the weight w_{ij} between v_i and v_j is positive or greater than a threshold. If $w_{ij} = 0$, or is less than the threshold, then v_i and v_j are not connected. A partition is found within the graph so that the edges between different groups have low weights and edges within groups have high weights (Von Luxburg, 2007).

The first step in the process is to calculate a *weighted adjacency matrix* for the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. This is essentially a matrix representing the similarity, or weight, between data points, or vertices. Some adjustment may be necessary to transform a similarity matrix to an adjacency matrix. For example, an adjacency matrix requires that $\omega_{ij} \geq 0$, so in the case of a correlation matrix, the correlations may need to be transformed so that they are all positive. Additionally, a vertex has no edge connecting to itself, so the diagonal of the adjacency matrix would consist of zeros not ones, as is

the case in a correlation matrix. Note, a particular weight need not be constrained $0 \leq \omega_{ij} \leq 1$, however the greater the weight, the greater the similarity between two vertices or data points. Note $\omega_{ij} = \omega_{ji}$.

Next the degree of each vertex, v_i, defined as $d_i = \sum_{j=1}^{n} w_{ij}$ (Von Luxburg, 2007), must be calculated. This is essentially the sum of all of the weights for all of the vertices connected to v_i by an edge. In practice this involves summing the rows of the adjacency matrix. A diagonal matrix is then formed (the *degree matrix)* with $d_{1,\ldots}d_n$ on the diagonal and 0 elsewhere.

Using the degree matrix and weighted adjacency matrix, a Graph Laplacian matrix can be calculated. The Graph Laplacian is the main requirement for spectral clustering. Graph Laplacian matrices are studied extensively in Spectral Graph Theory (Chung, 1997). A graph Laplacian can be un-normalized. This is calculated as $L = D - W$, where D is the degree matrix and W is the adjacency matrix. However, Von Luxburg (2007) recommends using the normalized Laplacian $L_n = I - D^{-1}W$.

Next the eigenvalues and eigenvectors are calculated for the Laplacian and arranged in order of increasing eigenvalue. Different methods have been proposed for selecting the number of eigenvectors to use for clustering. For example Shi & Malik (2000) use the k-means algorithm to cluster the first k eigenvectors into k clusters $C_{1,\ldots}C_k$. That is to say, the first k

eigenvectors corresponding to the smallest eigenvalues. Xiang & Gong (2008) proposed an approach which selects eigenvectors based on the information they provide in terms of being able to separate the data into clusters. Another approach is to use the second eigenvector (second smallest eigenvalue) which *gives a guaranteed approximation to the optimal cut* (Ng, Jordan, & Weiss, 2002), that is to say dividing $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ into two groups A and B, where $A \cup B = V$ and $A \cap B = \emptyset$ such that $Cut\ (A, B) = \sum_{i \in A, j \in B} w_{ij}$ is minimised. An example can be seen in Figure 9 where $Cut\ (A, B) = w_{34} + w_{15} = 0.3$.

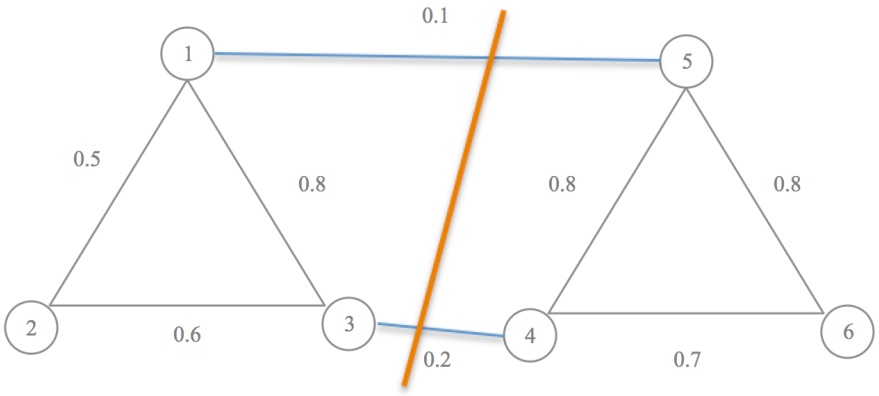


Figure 9 Example graph with six vertices showing the optimal cut.

Once the data have been transformed in this way, the clustering methodology is less important. Most approaches use k-means to cluster the points. There doesn't appear to be a theoretical reason for this,

although k-means is a well-known algorithm that is also empirically successful.

Of much more importance than the clustering method is the type of similarity measure used. Von Luxburg (2007) suggests it depends on the application area the data come from and that no general rule can be made. There are also different ways to construct the similarity graph. For example, the *ε-neighbourhood graph* connects all points whose distances are smaller than a certain threshold, the *k-nearest neighbour graph* connects v_i with v_j if v_j is within the *k* nearest neighbours of v_i (or vice versa) and the *fully connected graph* connects all points with positive weight w_{ij} weighting the edges by w_{ij}. Von Luxburg (2007) recommends using the k-nearest neighbour graph as it results in a sparser adjacency matrix.

6.2.1.1 Method

The similarity measure chosen was Kendall's Tau-b Correlation Coefficient due to the non-normal nature of the data and as preference over the Spearman rank correlation measure of association (Noether, 1986). Tau-b is typically applied to binary or ordinal data, so prior to running the analysis, all categorical variables were converted to dummy variables. Empty categories were then deleted.

A subset of the variables was selected using a Genetic Algorithm, to reduce the variable to observation ratio in the analysis, as there are a limited number of observations. Similar to the Genetic Algorithm described in section 5.2 a chromosome population was built consisting of 107 binary alleles, each representing a variable in the flourishing dataset CD_1995. Each allele had a probability of 0.2 of being assigned a 1 (select) otherwise it remained 0 (do not select). This meant that approximately 20 variables were chosen for each chromosome, to be used in the subsequent tournament. As described in section 5.3.2, two chromosomes were randomly selected at a time (without replacement) and a binary tournament took place.

The dataset was transposed such that the countries became the variables and the variables became the observations. The Kendall's Tau-b Correlation coefficient was calculated between each pair of a total of $p = 135$ countries and a $p \times p$ matrix, T, calculated for each subset of variables in the tournament, as defined by each chromosome. A typical element τ_{ij} is equal to the Kendall's Tau-b correlation between country i and country j. When visualised as a graph, the countries are the vertices and the adjusted correlation coefficients are the weights associated with the edges between them. The diagonal "ones" in the correlation matrix were replaced with "zeros" to create the adjacency matrix. That is $w_{ij} = \tau_{ij} - 1, where\ i = j$. The value of 1 was added to all remaining correlations to

ensure that all of the similarity measures were positive. That is $w_{ij} = \tau_{ij} + 1, where\ i \neq j$. This is the adjacency matrix of a fully connected graph and this approach was chosen for simplicity in regards to technical implementation.

The weighted degree matrix was calculated by summing the rows of the adjacency matrix, *W,* to give the degree, d_i of each vertex. That is, for $v_i \in V = \{v_1 \dots v_n\}$, $d_i = \sum_{j=1}^{n} w_{ij}$. The weighted degree matrix, *D,* is defined as the diagonal matrix with the degrees $d_1, \dots, d_n$ on the diagonal. Following which the normalised Laplacian, $L_n = I - D^{-1}W$, and its associated eigenvectors were calculated. The first eigenvector (with the smallest eigenvalue) was discarded, as this is always a vector of constant values approaching zero, when the graph is fully connected. For each member of the tournament, the second eigenvector was used to cluster the data into two groups (Algorithm I) as described by Ng et al. (2002). An alternative, identical algorithm (Algorithm II) used the second, third and fourth eigenvector to cluster the data into three groups as described by Xian & Gong (2008).

In addition to a manageable number of clusters and an even distribution of items among the clusters, of central importance in this case is the ability to profile the clusters using the data available in the original dataset. The distribution of the data and small sample size lends itself to a non-

parametric approach to measuring this ability. Therefore, the fitness function used to decide the winner of each tournament for the two-cluster (group) approach was the distribution free two-sample KS statistic. Kharoufeh, Goulias & Konstadinos G. (2002) use the Kolmogorov-Smirnov two sample test statistic to evaluate the size of the difference between two populations. This was the approach followed here, aiming to maximise 'differences' between the clusters. In this way the ability to describe the clusters using the variables, is maximised. The calculation works in the following way. For each member of the tournament, the original data were scored. In practice, this meant assigning each country to a cluster based on that particular solution. The original data were divided into two classes based on cluster membership. Then, for a given variable, the empirical distribution function was calculated for each class or cluster. The KS-test uses the maximum vertical deviation between the two curves (classes) as the statistic *D*. A visual example is shown in Figure 10. The KS-statistic was calculated, comparing the two classes, for each variable in the dataset. Larger values of *D* imply better separation between the clusters in regards to an individual variable. The average of the KS-statistic over all variables became the value of the fitness function. The winner of the tournament was the chromosome with the largest *D* with the aim of maximising the differences between clusters when they were profiled.

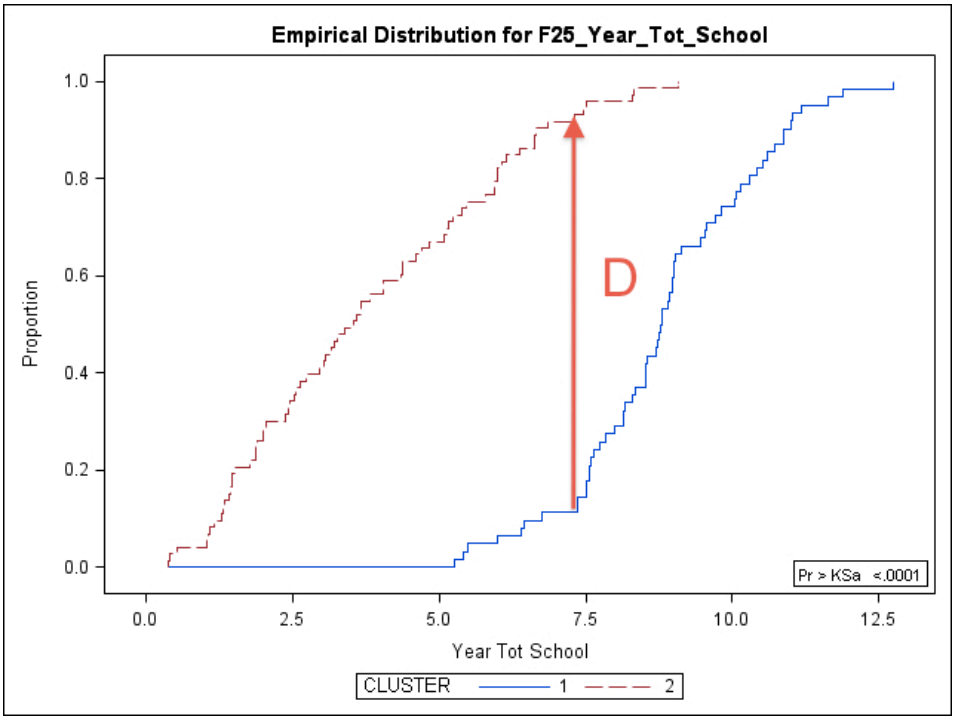


Figure 10: An example KS-test showing the statistic D as the maximum vertical distance between the empirical distribution functions of a particular variable – in this case years of total schooling for women over 25 – for two classes / clusters.

All tournaments in the original population proceeded in this way and the winners progressed onto crossover (varying the variables being used in the analysis) and mutation. The process repeated i.e. tournament, crossover, mutation, until the average value of the fitness function was stable for 20 generations.

The KS-Statistic can be used for two-sample comparison (Algorithm I), but for the three-cluster algorithm (Algorithm II), the fitness function needed to cope with multi-sample (cluster) solutions. There are many tests available for two-way samples (such as the previously described KS-Statistic) and a number available for multi-way samples (Hollander, Wolfe,

& Chicken, 1973). Of those available, the following were selected as suitable: The Kruskal-Wallis test, Brown-Mood test, Savage test and the Kolmogorov-Smirnov test statistics. It is possible to produce exact p-values for a number of the available tests (Narayanan & Watts, 1996), however here we are trying to ascertain the *general* picture of a cluster's profile over *all* the available profiling variables. For that reason the fitness measure involved computing the four statistics for each variable in the flourishing dataset CD_1995, for each clustering solution. In practical terms, similar to the approach used for the two cluster solution, this means assigning each country its relevant cluster based on the clustering solution in question, and then dividing CD_1995 into the different cluster populations and comparing the differences between the 'populations' for each variable. The mean was then calculated for each statistic over all variables and then the sum of the means was stored as that clustering solution's fitness. A visual representation of the multi-way fitness function can be found in Figure 11. Again, the winner of each tournament was the chromosome with the largest fitness value.

Cluster	Countries in Cluster	Proportion with No Education	Proportion of population aged 0-14	Proportion of women in total Population	Average Years of Primary School
1	Norway, Denmark, The Netherlands etc.	w_1	x_1	y_1	z_1
2	NZ, UK, Australia, Canada etc.	w_2	x_2	y_2	z_2
3	Argentina, Chile, Brazil etc.	w_3	x_3	y_3	z_3
4	Zimbabwe, Egypt, South Africa etc.	w_4	x_4	y_4	z_4

Statistics to compare cluster differences (by variable)	Kruskal-Wallis	KW_1	KW_2	KW_3	KW_4	$\overline{KW} = \frac{\sum_{t=1}^{4} KW_t}{4}$
	Brown-Mood	BM_1	BM_2	BM_3	BM_4	$\overline{BM}$
	Savage	S_1	S_2	S_3	S_4	$\overline{S}$
	Komogorov-Smirnov	KS_1	KS_2	KS_3	KS_4	$\overline{KS}$

Fitness Function	$\overline{KW} + \overline{BM} + \overline{S} + \overline{KS}$

Figure 11: The multi-way fitness function calculation method.

6.2.1.2 Results

6.2.1.2.1 Two Cluster Solution – Algorithm I

For Algorithm I, the 17 variables selected by the Genetic Algorithm for the optimal solution, that is the solution that produced the best descriptive separation between the clusters, are found in Table 13.

Variable Name	Variable Description	Can 1	Can 2
DPI_SYSTEM2	Regime type (0) Direct presidential (1) Strong president elected by assembly (2) Parliamentary	0.1399	-0.3473
F15_PRIM_TOT	Percentage of population Female 15+ whose highlest level of education is primary	0.1212	0.2009
F25_NO_SCHOOLING	Percentage of population Female 25+ with no schooling	1.2339	-0.1875
F25_YEAR_PRIM_SCHOOL	Females 25+ average years of primary schooling	0.3180	1.8188
F25_YEAR_TOT_SCHOOL	Females 25+ average years of total schooling	0.0185	-3.9986
FE_PLURAL	Plurality group: population share of the largest group	-0.1950	0.0620
FM_AST_DOMS_CN	Net domestic credit (current LCU)	0.0220	0.1095
FPR_25N29	Female Labour Force Participation rates Aged 25 to 29	0.2037	-0.1312
FPR_55N59	Female Labour Force Participation rates Aged 55 to 59	0.7904	0.7235
FPR_65P	Female Labour Force Participation rates Aged 65+	0.1833	-0.5700
LP_PROTMG80	Protestants as % of population in 1980	-0.2125	-0.1028
MF15_SEC_TOT	Percentage of population 15+ whose highlest level of education is secondary	-0.3251	1.8772
NY_GDP_PCAP_CD	GDP per capita (current US$)	-0.0204	-0.3479
SE_ENR_PRIM_FM_ZS	Ratio of female to male primary enrollment (%)	-0.1063	-0.0747
SE_PRE_ENRR	School enrollment, preprimary (% gross)	-0.0467	0.2809
SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)	-0.3602	0.0453
SH_XPD_PCAP_PP_KD	Health expenditure per capita, PPP (constant 2005 international $)	-0.1131	0.9538
SP_POP_TOTL	Population, total	0.0415	-0.1610

Table 13: The 17 variables, and their associated total-sample standardized canonical coefficients, selected by the Genetic Algorithm for the two-cluster spectral solution.

In this solution there are a noticeably large number of variables focused on the education and employment of women. GDP and health expenditure per capita are also included. The results of cluster analysis are often displayed using canonical discriminant analysis. This analysis method attempts to find linear combinations of the independent variables $\boldsymbol{x}$, in this case the variables in the CD_1995 dataset, that provide maximum separation between the groups $\boldsymbol{y}$, in this case dummy variables coded to represent the number of clusters. (McLachlan, 2004)

Vectors a and b are sought such that the random variables $U = a'\boldsymbol{x}$ and $V = b'\boldsymbol{y}$ maximize the correlation $\rho = corr(a'\boldsymbol{x}, b'\boldsymbol{y})$. The random variables, U and V, are the first pair of canonical variables. Subsequent vectors are sought, to give the second pair of canonical variables, subject to the constraint that they are un-correlated with the first pair of canonical variables. This procedure can be continued up to $min\{m, n\}$ times, where in this case m will be the number of clusters. In canonical discriminant analysis the first member of each pair of canonical variables i.e. $a'\boldsymbol{x}$, are new variables that provide maximum separation between the groups. The values in a are the coefficient weights or loadings on the original variables.

This procedure was applied, to provide the initial visualisation found in Figure 12. The first canonical variable, the x-axis in Figure 12, has a

canonical correlation of 0.89. That is to say 0.89 is the greatest multiple correlation with the cluster membership that can be achieved by using a linear combination of the variables in Table 13. An analysis of the canonical structure of this dimension suggests that moving from left to right on this axis would tend to bring increased amounts of older women in the workforce and adult women with no schooling. It would also tend to bring lower amounts of measles immunization, health expenditure per capita – PPP and lower amounts of secondary schooling for both men and women. Multivariate normality is not strictly required for this type of analysis, but it is desirable, and data transformation is recommended if necessary (Hair, Anderson, Tatham, & Black, 1995). For this reason, this visualisation is an initial approach only.

6.2.1.2.2 Three Cluster solution – Algorithm II

For Algorithm II the 26 variables chosen by the Genetic Algorithm as the optimal choice for the 3-cluster spectral clustering solution are shown in Table 14.

Variable Name	Variable Description	Can 1	Can 2
_eco__agri_area_	Agricultural area square kilometres	-0.3170	-0.0156
_eco__terr_prote	Protected areas - square kilometres	0.0557	-0.2724
_gen__pop_rural_	Population residing in rural areas 000's	1.9979	2.7159
_gen__pop_total_	Population: de facto population in a country, area or region as of 1 July of the year indicated (000's)	-4.4478	-2.1193
_res__cereals_yi	Actual cereals yielded hectograms per hectare	0.0473	0.0636
DT_ODA_ALLD_KD	Net official development assistance and official aid received (constant 2008 US$)	-0.0261	0.2283
F15_Prim_Tot	Percentage of population Female 15+ whose highlest level of education is primary	-0.2016	0.0591
F25_Year_Prim_School	Females 25+ average years of primary schooling	0.4315	-0.1137
fe_plural	Plurality group: population share of the largest group	0.0076	0.0357
fh_cl	Civil liberties: 1 (most free) and 7 (least free)	-0.1686	-0.2883
FPR_55N59	Female Labour Force Participation rates Aged 55 to 59	-0.3606	-0.1489
ht_regtype11	Regime type monarchy	0.9069	0.5574
ht_regtype14	Regime type multi-party	-0.0765	-0.3878
MF15_Pop_N_000s	Number of Population 15+	2.9583	1.2308
MF15_Year_Tert_School	Population 15+ average years of tertiary schooling	-0.1564	0.1624
MFPR_40N44	Male and Female Labour Force Participation rates Aged 40 to 44	-0.3544	0.4624
MFPR_50N54	Male and Female Labour Force Participation rates Aged 50 to 54	0.4721	0.5513
NY_ADJ_DKAP_GN_ZS	Adjusted savings: consumption of fixed capital (% of GNI)	0.3348	-0.1441
SE_PRE_ENRR	School enrollment, preprimary (% gross)	0.2979	0.1173
SE_PRM_AGES	Primary school starting age (years)	-0.2217	-0.0817
SH_XPD_PCAP_PP_KD	Health expenditure per capita, PPP (constant 2005 international $)	0.0287	0.5150
SH_XPD_PRIV_ZS	Health expenditure, private (% of GDP)	0.0538	-0.1162
SH_XPD_PUBL_GX_ZS	Health expenditure, public (% of government expenditure)	0.2165	0.1959
SL_TLF_TOTL_IN	Labor force, total	-0.4279	-1.9318
SP_DYN_CDRT_IN	Death rate, crude (per 1,000 people)	-0.0011	-0.3888
SP_DYN_LE00_FE_IN	Life expectancy at birth, female (years)	0.6456	-1.5709

Table 14: The 26 variables, and their associated total-sample standardized canonical coefficients, selected by the Genetic Algorithm for the three-cluster spectral solution.

There are more health related variables in this solution, but still a large focus on education. Of interest is the regime type and civil liberties variable. A canonical discriminant analysis was again carried out and a plot of the first two canonical variables can be found in Figure 13.

The first canonical variable has a canonical correlation of 0.9 and appears to separate cluster two from the remaining cluster one and three, and the second canonical variable has a canonical correlation of 0.8 and appears to separate cluster three from clusters one and two. Moving to the right of the first variable, the x-axis, would bring generally increasing: years of primary school for adult women, pre-primary school enrolment, life expectancy at birth for women, health expenditure as a proportion of government spending, adjusted savings: consumption of fixed capital (%

of GNI), and smaller populations. Moving from bottom to top up the y axis would tend to bring increased workforce participation for men and women over 40 and particularly over 50, smaller populations and more people residing in rural areas. There will likely be variations to these tendencies for countries along the spectrum where for example a country may not follow all of the tendencies expected.

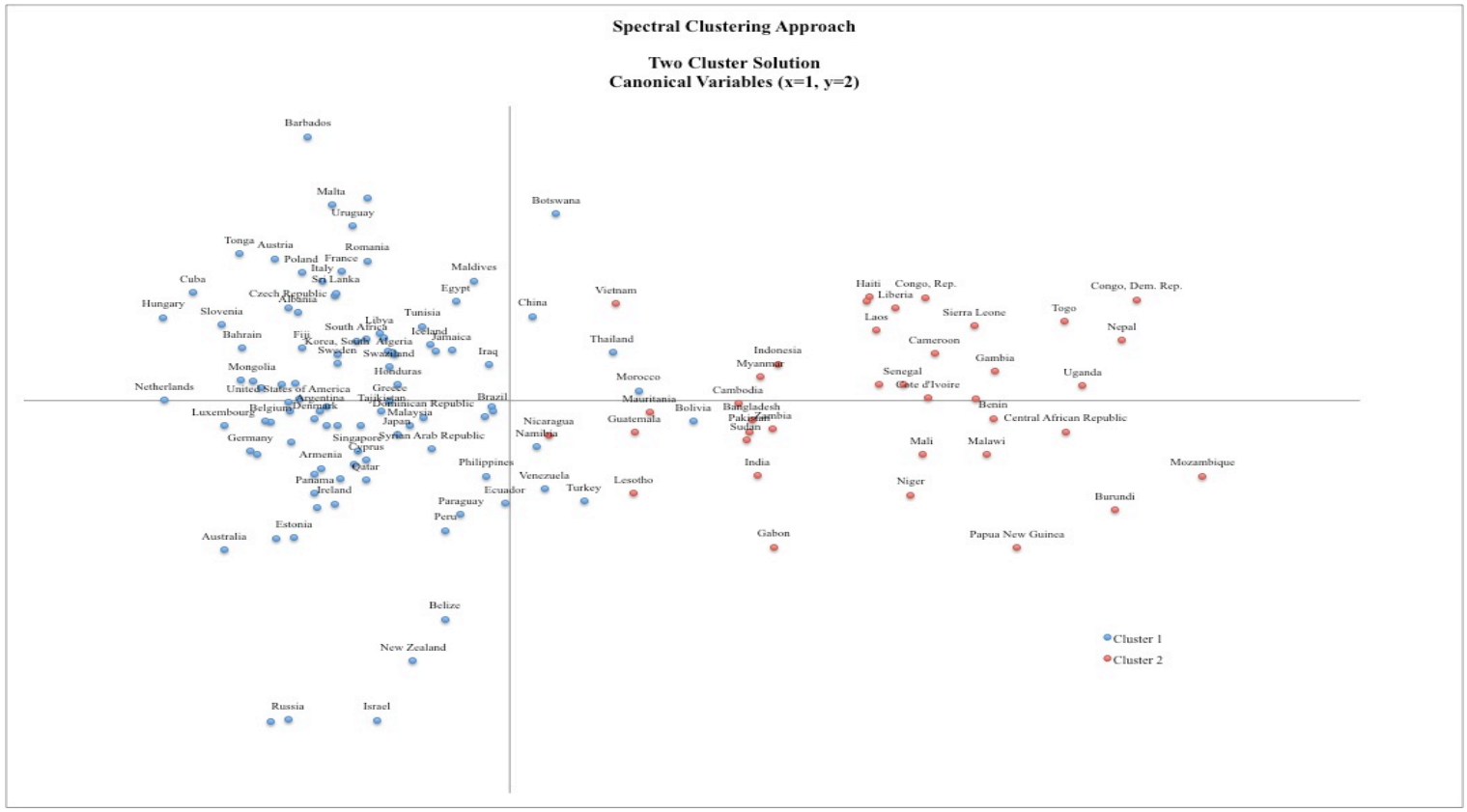


Figure 12: A plot of the first two components from a canonical analysis on the 2-cluster spectral method.

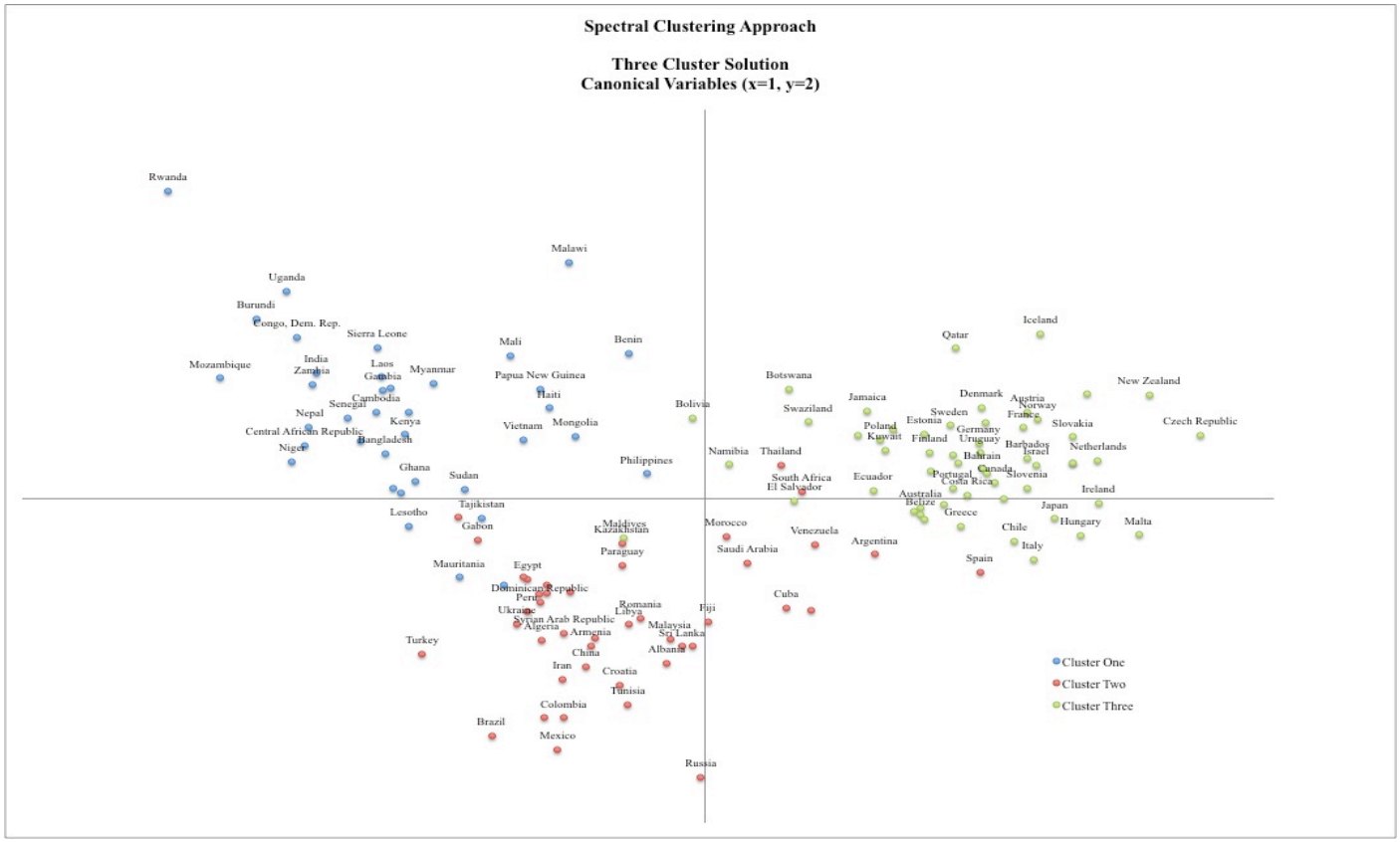


Figure 13: A plot of the first two components from a canonical analysis on the 3-cluster spectral method.

6.2.1.3 Summary

A Genetic Algorithm was constructed to select the 'best' variables from the flourishing dataset CD_1995 for inclusion in a Spectral Clustering analysis. Two Spectral Clustering approaches were used – Algorithm I used the 'optimal cut' second eigenvector (second smallest eigenvalue) to produce two clusters or groups of the countries available, and a second approach (Algorithm II) used the first k=3 eigenvectors to produce three clusters or groups. When plotted using canonical dimensions, the three-cluster solution gave better separation between the countries, and will therefore be investigated further.

Both solutions were visualised using variables produced by a canonical discriminant analysis. These produced analyses centred largely on education – in particular the education of women, and healthcare. However, CD_1995 is not a multivariate normal dataset and because of this, and the potential assumption violations, the visualisation and profiling requires further work.

6.2.2 Genetic Algorithm For Clustering

The previous sections described two applications of spectral clustering analysis. Jain (1999) describe a further assortment of available algorithms including hierarchical, partitional, mixture resolving, nearest neighbour,

fuzzy and evolutionary algorithms. They also point out that having a large number of algorithms available can cause a dilemma for a user, particularly one short on expertise or time, who is trying to choose the best approach for any particular problem. An analyst may consistently choose to use a familiar or favourite algorithm e.g. k-means, which may not be the best choice for all datasets.

Genetic Algorithms have been used to directly cluster data, as seen for example in Ujjwal & Sanghamitra (2000). However at the time of writing, I have not seen a Genetic Algorithm used to choose between a selection of available algorithms within a software package. In section 5.2, a Genetic Algorithm was used to choose the variables to be included and method to be used (from either forward backward or stepwise regression), to perform a regression analysis. The root mean squared error was used to decide which group of variables and method of regression was the 'best' in terms of predicting the dependent variable 'life satisfaction'.

In this application, a Genetic Algorithm was designed written and used to select the most appropriate clustering method from a range of hierarchical agglomerative methods. The choice of methods was limited in this way, for practical purposes, and with a view to the algorithm being transferrable to other applications, such as a data analyst investigating data in a commercial setting. As there are a large number of variables relative to the number of observations, the algorithm was also used to select the 'best'

subset of variables from those available in the CD_1995 dataset, to then be used in the clustering analysis.

6.2.2.1 Method

An initial sample population was generated to represent the potential solution space. Similar to section 5.3.2, the population was made up of chromosomes, this time consisting of 111 alleles. The first 107 alleles in each chromosome represented the variables available in the flourishing dataset CD_1995. They were binary alleles, randomly assigned 1 with probability, Pr=0.2 or 0 with probability, Pr = 0.8. This restricted the variable to observation ratio in the cluster analysis to approximately $1/5^{th}$ (Osborne & Costello, 2004). Allele 108 and 109 were also binary and represented outlier treatment and standardisation of the original data. Both of these alleles could be 0 or 1 with equal probability. Allele 110 controlled the type of standardisation, if standardisation had been selected for. This allele took on a value from 1 to 4 with equal probability. These represented the standardisation methods outlined in Table 15. The first three methods were chosen for their various properties with regards to robustness in clustering small data with outliers (Iglewicz, 1983) (Goodall, 1983).

Value	Method	Location	Scale
1	AGK	Mean	AGK estimate (ACECLUS)
2	MAD	Median	median absolute deviation from median
3	IQR	Median	interquartile range

4	STD	Mean	standard deviation

Table 15: The standardisation methods used in the Genetic Algorithm approach to clustering.

The final allele took on a value of 1 to 8 where each value represented a clustering method. A range of methods were considered, although some were considered inappropriate including 'complete linkage' (Sorensen, 1948) which is distorted by outliers; the 'EML method' (Symons, 1981) as it requires multivariate normal data; and Ward's Minimum-Variance Method (Ward, 1963) as it also requires multivariate normal data. The value and associated method can be found in Table 16. Note the average method appears twice as it replaced the Uniform method that was, during testing, also found to be unsuitable in this case. This meant seven unique methods with one of those having twice the probability of those remaining of being selected for testing. It was of interest to observe if this would weight the outcome, however it did not appear so as the average method was not selected as optimal.

Value	Method
1	Average
2	Centroid
3	k-nearest neighbour
4	Average
5	Flexible
6	McQuitty's
7	Median
8	Two-stage

Table 16: The clustering methods used in the Genetic Algorithm approach to clustering.

Because cluster analysis works best with binary or continuous data, prior to running the analysis, and similar to 6.2.1.1, any categorical variables that had been selected for analysis were converted to dummy variables (Reiss, 2010). The dummy variables (associated with classes within the categorical variables) containing less than 30 members were then deleted, as dummy variables containing less caused an error in algorithm processing. Following the same process as outlined in section 5.3.2, the sample population of chromosomes was randomly sampled into groups of two for a binary tournament. A cluster analysis was run for each chromosome, with the conditions as determined by the relevant alleles. The 3% most distant outlying observations were trimmed lightly for all analyses, but if the allele controlling outlier treatment was positive, then the 5% most distant outlying observations were trimmed. The results were automatically analysed in terms of traditional clustering statistics - the cubic clustering criterion (ccc), pseudo F and pseudo t^2, to determine an appropriate number of clusters. These statistics often give a number of potential results, therefore up to three possible results, i.e. potential number of clusters, were stored. These results were determined in the following way. Firstly, for a potential solution containing a particular number of clusters, $x_i \in X = \{x_1 \dots x_8\}$, a dataset consisting of a subset of the hierarchical structure was obtained containing *X* and the associated ccc, $c_i \in C = \{c_1 \dots c_8\}$, pseudo F, $f_i \in F = \{f_1 \dots f_8\}$, and pseudo t^2,

$t_i \in T = \{t_1 \dots t_8\}$, values for each x_i. These were then processed as following:

$$a_i = f_{i-1}, b_i = f_{i-2}$$

$$y_i = \begin{cases} a_i - b_i & if\ (a_i > f_i \wedge b_i < a_i) \\ 0 & otherwise \end{cases}$$

$$z_i = t_i - t_{i-1}$$

The elements (values) of the vectors (columns) $X_{8,1} C_{8,1}, Y_{8,1}$ and $Z_{8,1}$ were combined into a matrix (dataset) $S_{8,4}$ and replaced by the values of their ranks. $S_{8,4}$ was then processed as follows:

$$if\ y_i = j \rightarrow m_i = x_{i-1}, \quad for\ i,j = \{1,2,3\}$$

$$if\ z_i = j \rightarrow n_i = x_{i-1}, \quad for\ i,j = \{1,2,3\}$$

$$if\ c_i = j \rightarrow o_i = x_i, \quad for\ i,j = \{1,2,3\}$$

It is important to note that although each solution was limited to three possible results, as above, some solutions had less than this. This did not impact processing, although the algorithm was programmed to discard the solution if there were no potential results. The values of m_i, n_i and o_i were then combined such that

$$S2 = [m_1, n_1, o_1, m_2, n_2, o_2, m_3, n_3, o_3]$$

for all existing m_i, n_i and o_i. *S2,* a potential 1 x 9 matrix, was then transposed, such that the original values of x_i became the column vector.

$$S3 = S2'$$

S3 was then summarised to determine how many times each unique value appeared. Only values that appeared more than once were kept. The 'best' solution was determined to be the 'value' or number of clusters that appeared the most often i.e. was picked most frequently by the ccc, pseudo F and pseudo t^2, and the number of times the value appeared. These two components were called 'ccc fit' and 'agree fit' respectively. The code associated with this automatic analysis of the clustering statistics, can be found in Appendix 4.

These two components formed the first part of the fitness function, which continued from here with two approaches.

6.2.2.1.1 Fitness Function Approach One

As mentioned, if there was no acceptable solution with less than 8 clusters (from the three possible results), then the fitness was set to 0 and that chromosome was discarded. Eight was chosen as the limit for manageability.

The algorithm then continued with the 'best' of the possible results (as described in the previous section), for each member of the tournament,

storing both the number of clusters (referred to as ccc fit) and the 'level of agreement' (either two, or three statistics agreed). A visual example can be found in Figure 14.

Chromosome	Statistics to determine Number of Clusters			Result
	ccc	Pseudo F	pseudo t^2	
1	5 6	3 5	2 5	2 possible solutions Agree Fit = 3 ccc fit = 5
2	4 5 6	3 4 7	4 6 8	3 possible solutions Agree Fit = 3 ccc fit = 4

Figure 14: An example showing programmed method of cluster number detection.

At this point the frequencies within the clusters were checked using the "ccc fit" solution. This was to ensure that only solutions with reasonably equal counts of countries within their clusters were able to continue. This is because the research aims to find a way of measuring flourishing that will be stable across time. Any solution where the proportion of data in the biggest cluster was more than 30% larger than the proportion of countries in the smallest cluster was discarded. The 30% threshold was selected through trial runs of the algorithm. The range, that is the difference between the smallest and largest cluster proportions, was stored as a further component of the fitness function (the smaller the range, the better).

There are a number of items that are important to the fitness of this cluster analysis. As well as traditional methods and the distribution amongst the 'clusters', as mentioned in 6.2.1.1, the way the solution 'profiles' is also

very important. It is important that between-cluster differences are maximised. As also mentioned in 6.2.1.1, Kernel Density Estimation provides distribution free methods for evaluating differences between populations. In the previous application the KS-Statistic was used to compare two populations, and as is the case in this application, the average of The Kruskal-Wallis test, Brown-Mood test, Savage test and the Kolmogorov-Smirnov test statistics were used to examine the differences between multiple populations i.e. testing how the distribution of a variable varies over each cluster.

All previously mentioned parts of the fitness function were then combined to produce the fitness function for each competitor in the tournament. This was done using a formula:

$$\mathcal{F} = \mathcal{A} - \mathcal{C} - \mathcal{R} + (1000 * \mathcal{D})$$

where $\mathcal{F}$ is the fitness of the competitor, $\mathcal{A}$ is the agreement level between the cubic clustering criterion, pseudo F and pseudo t^2, as described in 6.2.2.1 and $\mathcal{C}$ is the number of clusters selected by the cubic clustering criterion, pseudo F and pseudo t^2 method as described in 6.2.2.1. $\mathcal{R}$ is the difference between the proportion of countries within the largest cluster, and the proportion of countries within the smallest cluster, and was restricted to values less than 30. $\mathcal{D}$ is the sum of the average values of the distribution test statistics. $\mathcal{D}$ was scaled up by a factor of 1,000 to bring it

in line with the scale of the other members. In this case, this scaling is because there is no preference regarding the importance of each component of the fitness function, however, in another application it may be desirable to place more or less importance on individual components of the fitness function.

6.2.2.1.2 Fitness Function Approach Two

The approach in 6.2.2.1.1 involved scaling a component of the fitness factor to bring it in line with the other components. This is because, for this application, there is no preference to place more importance on any individual component of the fitness function. However, in another application, this ability may be useful. For this research, where no weighting is required for the components of the fitness function, an alternative approach to calculating fitness was trialled. This involved ranking each component of the fitness function with respect to the other members of the trial (other chromosomes in the trial). The rank of each component was then combined to create the overall fitness for each chromosome. This was done by separating the components of the fitness function, for all members of the trial, into two groups. The first group contained the values $\mathcal{A}$ and $\mathcal{D}$ as described in 6.2.2.1.1, and the associated chromosome identifier for each chromosome within the trial. Each value of $\mathcal{A}$ and $\mathcal{D}$ was replaced by the value of its respective rank within the group. For this part of the fitness function, greater values received greater

ranks. The second group contained the values $\mathcal{R}$ and $\mathcal{C}$ as described in 6.2.2.1.1, and the associated chromosome identifier. Each value of $\mathcal{R}$ and $\mathcal{C}$ was replaced by the value of its respective rank, however this time the values were ranked from largest to smallest. In this way smaller values received greater ranks. The fitness function, where $\mathcal{F}$ is again the fitness of the competitor, then became:

$$\mathcal{F} = \mathcal{A} + \mathcal{D} + \mathcal{R} + \mathcal{C}$$

For both approach one and two, and similar to the method used in section 5.2, the chromosome with the greater fitness of the two then proceeded to the next stage of crossover and mutation. For approach one, 50 tournaments were run on the first loop and from the 50 fittest, 30 proceeded to crossover and mutation. For approach two, the population size and tournament count were parameterised. The final trial settled on an initial population of 300 chromosomes, resulting in 150 tournaments. The 'fittest' 100 then proceeded to crossover and mutation. The top chromosome was stored as 'elite' and the remainder resampled into groups of two. At this point crossover occurred at a random point between the two members of each group. The allele chosen was always within the variable list such that all alleles following (i.e. any remaining variables, the outlier, standardisation and clustering treatment) were always exchanged. Mutation occurred in 0.1% of cases at the crossover point. That is, the value of the allele was reversed from '0' to '1' or from '1' to '0'.

6.2.2.2 Results

6.2.2.2.1 Fitness Function Approach One

The 21 variables selected by the Genetic Algorithm for the optimal solution, that is the solution that produced the best descriptive separation between the clusters, are found in Table 17.

Variable Name	Variable Description	Can 1	Can 2
_gen__pop_female	Defacto population as of 1 July of year indicated - female	-1.3758	-7.0182
_gen__pop_total_	Population: de facto population in a country as of 1 July of the year indicated (000s)	-1.5096	-5.7185
_res__cereals_yi	Actual cereals yielded hectograms per hectare	0.3229	0.3272
affected	People affected by natural disasters	0.0002	0.2361
BX_KLT_DINV_CD_WD	Foreign direct investment, net inflows (BoP, current US$)	0.1889	0.3583
DT_ODA_ALLD_KD	Net official development assistance and official aid received (constant 2008 US$)	0.0920	0.1519
F15_Prim_Tot	Percentage of population Female 15+ whose highest level of education is primary	-0.1874	0.3914
F25_Year_Prim_School	Females 25+ average years of primary schooling	0.0185	-1.0018
lp_lat_abst	Latitude	0.3016	-0.7170
MFPR_25N29	Male and Female Labour Force Participation rates Aged 25 to 29	-0.1189	0.2545
MFPR_45N49	Male and Female Labour Force Participation rates Aged 45 to 49	0.1926	-0.6297
MPR_65P	Male Labour Force Participation rates Aged 65+	-0.5504	0.5202
NY_GDP_PCAP_KD	GDP per capita (constant 2000 US$)	2.1253	1.5993
SE_PRM_DURS	Primary education, duration (years)	-0.1708	0.9973
SH_IMM_IDPT	Immunization, DPT (% of children ages 12-23 months)	0.0852	-0.2037
SH_XPD_OOPC_TO_ZS	Out-of-pocket health expenditure (% of total expenditure on health)	-0.1935	0.0689
SH_XPD_PUBL_GX_ZS	Health expenditure, public (% of government expenditure)	-0.0225	-0.1286
SL_TLF_TOTL_IN	Labour force, total	-0.0754	-1.5393
SP_DYN_CBRT_IN	Birth rate, crude (per 1,000 people)	0.5611	0.0934
SP_DYN_LE00_FE_IN	Life expectancy at birth, female (years)	0.5248	0.0707
SP_URB_TOTL_IN_ZS	Urban population (% of total)	0.3810	-0.1238

Table 17: The 21 variables, and their associated total-sample standardized canonical coefficients, selected by Fitness Function Approach 1.

These variables have a similar focus to the variables selected in the three-cluster spectral solution, but with less focus on education. Included are environmental variables, such as agricultural yields, variables regarding health, labour force participation and economic wealth. The algorithm selected the Flexible Beta clustering Algorithm, with additional outlier restriction and standardisation using the Standard methodology. The cluster distribution produced by this solution, showing the number of countries within each cluster, can be found in Table 18, showing a

relatively even spread of countries among seven clusters. Seven countries were classed as outliers, due to the tight outlier restriction, and were unable to be assigned a cluster.

Cluster	Count	Proportion %
1	20	16
2	20	16
3	21	16
4	19	15
5	21	16
6	12	9
7	15	12
None Assigned	7	.

Table 18: The number of countries within each cluster from the solution produced by Fitness Function Approach 1.

As in Section 6.2.1.2, a canonical discriminant analysis was carried out to provide the initial visualisation found in Figure 15. There were four significant canonical variables. The first canonical variable, the x-axis in Figure 15, has a canonical correlation of 0.96. This is the greatest multiple correlation with the cluster membership that can be achieved by using a linear combination of the variables in Table 17. The second canonical variable, with a canonical correlation of 0.95, can be interpreted similarly, subject to the constraint that it cannot be correlated with the first canonical variable. An analysis of the canonical coefficients of the first dimension suggests that moving from left to right on this axis, one could expect to see a generally increasing absolute value of the latitude of the capital city. This is the strongest component of the linear combination. There may

also tend to be a slight decrease in people who are employed over age 65 and average years of primary education.

Moving from bottom to top on the second dimension, you could expect to see generally increasing primary education with a slightly decreasing average years of primary school for women over 25. These statements would not always be the case for an individual country, due to the variation within each country's data and the linear combination of the variables. For example a very large result in one area may counteract a very small result in another meaning that country may not conform to the 'generally' statement.

Of note in Figure 15, is that although there is separation of some clusters in two dimensions, other clusters are less distinct. Also, to aid visualisation/readability, only some of the data points are labelled.

6.2.2.2.2 Fitness Function Approach Two

The 14 variables selected by the Genetic Algorithm for the optimal solution are found in Table 19.

Variable Name	Variable Description	Can 1	Can 2
_gen__mobile_pho	Number of phones per 100 inhabitants	21.9985	1.4834
_res__cereals_ha	Actual cereals harvested sq kilometres	0.0697	0.0788
F25_No_Schooling	Percentage of population Female 25+ with no schooling	0.1833	0.0232
IT_CEL_SETS_P2	Mobile cellular subscriptions (per 100 people)	-21.7405	-1.0915
MF25_Year_Prim_School	Population 25+ Average years of primary schooling	0.4921	-0.4781
MFPR_30N34	Labour Force Participation rates Aged 30 to 34	0.1667	-1.0560
NY_ADJ_DCO2_GN_ZS	Adjusted savings: carbon dioxide damage (% of GNI)	-0.2350	0.1762
NY_ADJ_DKAP_GN_ZS	Adjusted savings: consumption of fixed capital (% of GNI)	0.1699	0.1671
NY_GDP_MKTP_PP_KD	GDP, PPP (constant 2005 international $)	0.1671	0.1720
NY_GDP_PCAP_CD	GDP per capita (current US$)	0.8989	1.2367
SH_XPD_PUBL_GX_ZS	Health expenditure, public (% of government expenditure)	0.0827	0.3178
SL_TLF_CACT_ZS	Labour participation rate, % of total population ages 15+	0.1316	0.2035
SP_POP_0014_TO_ZS	Population ages 0-14 (% of total)Ê	0.5218	1.9446
SP_POP_65UP_TO_ZS	Population ages 65 and above (% of total)	2.4240	0.8764

Table 19: The 14 variables, and their associated total-sample standardized canonical coefficients, selected by Fitness Function Approach Two.

Similarly to the first genetic algorithm approach, there are a variety of variables in this selection, covering mobile phone use, education, CO_2 emissions, health, labour force participation and populations stratified by age. This algorithm also selected the Flexible Beta Clustering Algorithm and the Standard standardisation method, however it did not select for extra outlier restriction. The result was a six-cluster solution, and the distribution of countries among clusters can be found in Table 20.

Cluster	Count	Proportion %
1	22	17
2	23	18
3	20	15
4	29	22
5	17	13
6	19	15
None Assigned	5	.

Table 20: The number of countries within each cluster from the solution produced by Fitness Function Approach 2.

A canonical discriminant analysis again produced five statistically significant variables, indicating the variation in the results cannot be fully visualised using two dimensions. The canonical correlations of the first two variables are 0.95 and 0.83 respectively and a plot of these two

variables can be found in Figure 16. Moving to the right across the x-axis (first variable) you could expect to see more phones per 100 inhabitants but conversely less cellular subscriptions per 100 people, slightly decreasing carbon dioxide damage as a % of GNI (adjusted savings) and a slightly increasing population aged 65+ as a % of total population. Moving from the bottom up to the top on the y-axis (second variable), you could expect to see slightly increasing carbon dioxide damage as a % of GNI, and a slightly decreasing number of years at primary school. Again there is some good cluster separation over two dimensions, but not all can be separated in this way suggesting this two dimensional solution, and the interpretation of the dimensions, is insufficient.

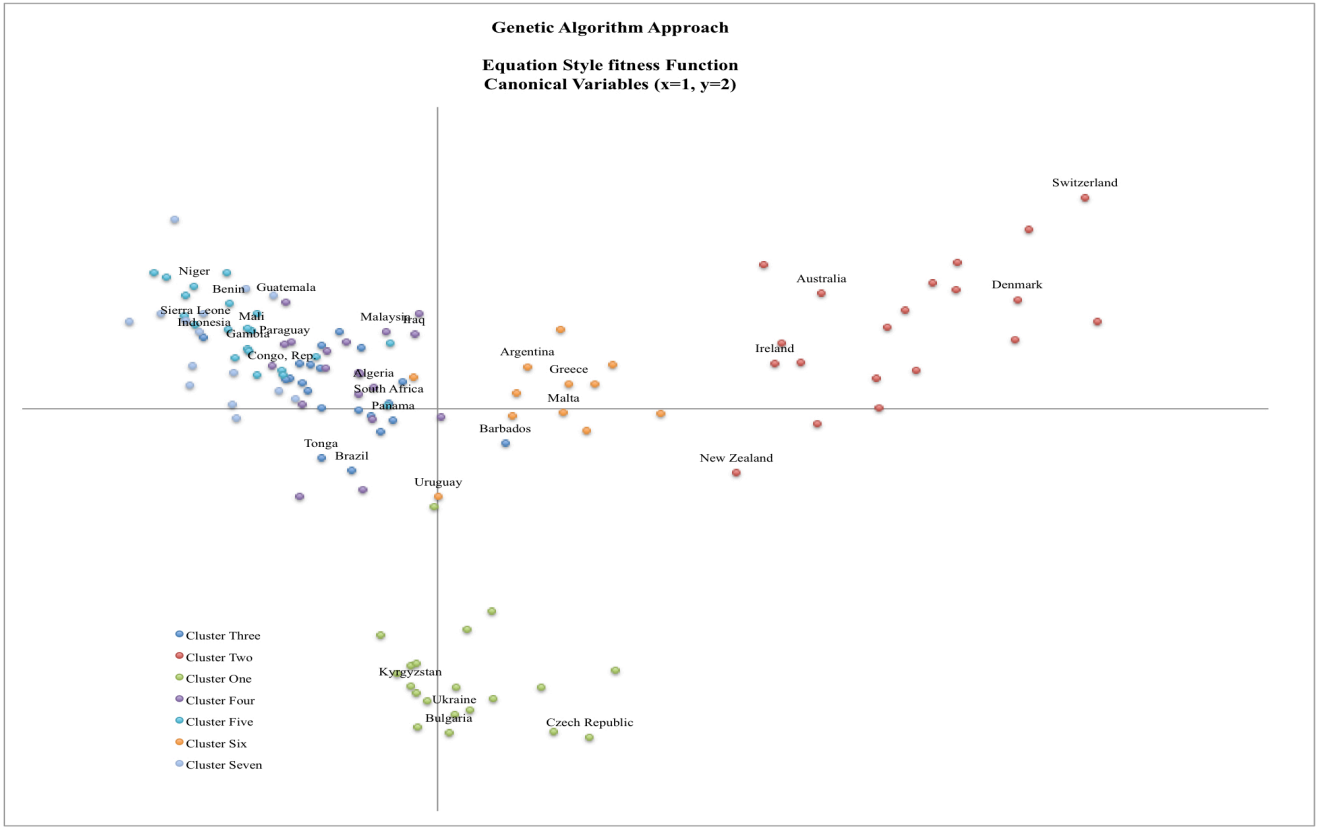


Figure 15: A plot of the first two canonical variables from a Canonical Analysis of the solution produced by Fitness Function Approach 1.

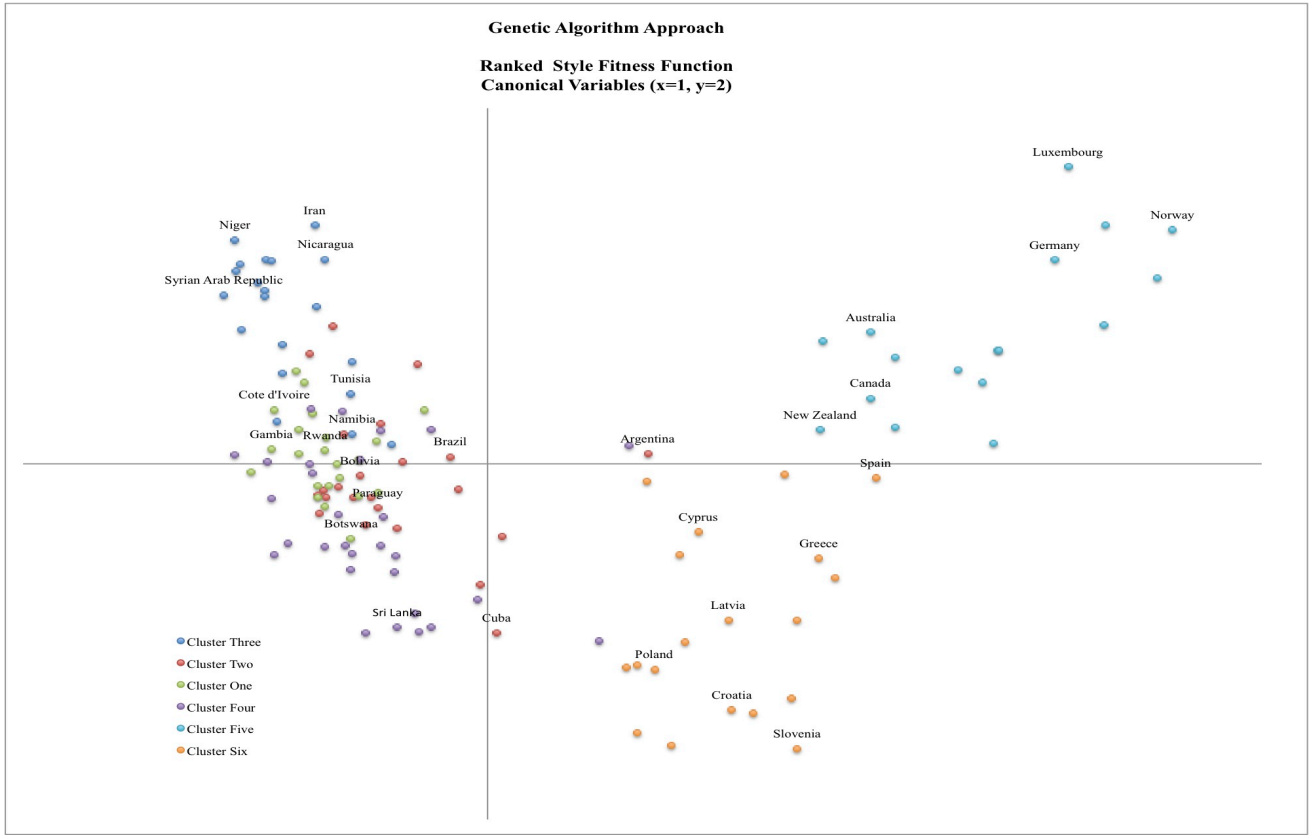


Figure 16: A plot of the first two canonical variables from a Canonical Analysis of the solution produced by Fitness Function Approach 2.

6.2.2.3 Summary

A genetic algorithm was used to choose the 'optimal' clustering method from a range. Two approaches to the fitness function were used, the first one combined agreement between the statistics used to determine number of clusters, the resultant proposed number of clusters, distribution within the clusters and the solution's ability to provide interpretable differentiation and profiling of the clusters. The measures were scaled to ensure they had similar weight in the fitness function. The second approach ranked each of these measures, and then combined the ranks to give the fitness function. Both fitness functions produced results with a reasonable number of clusters and a reasonable distribution of countries between clusters. The ability to restrict the number and size of the clusters in this way could have many useful applications. The components of the fitness function could also be easily varied. Neither of the solutions were satisfactorily visualised using the first two canonical variables and further work is required from here to profile and visualise the clusters. I will progress with the ranked version of the fitness function because in this case there is no requirement to apply different weights to individual members of the fitness function.

6.2.3 Comparative Clustering Technique: K-means

K-means is a popular clustering algorithm (Wagstaff, 2001). As a comparative technique a K-means clustering analysis was carried out on the full set of data i.e. all 107 variables. The data were standardised to mean 0 and variance 1. The variable to observation ratio is high at almost 1:1, but this highlights the need for an easy way to choose the 'best' variables for an individual problem. The maximum number of clusters for k-means clustering needs to be pre-set, so for comparison to the Genetic Algorithm approach a value of six was chosen. The analysis produced a very uneven distribution of countries between the clusters, as shown in Table 21. Reducing the data to an arbitrary selection of 20 variables still produced a similar outcome.

Cluster	Count	Proportion %
1	3	2
2	1	1
3	1	1
4	2	1
5	126	93
6	2	1

Table 21: The number of countries within each cluster from the solution produced by K-Means clustering.

K-means works by choosing *k* random data points from the dataset to serve as the centres for *k* clusters. Then the distance from each of the other data points to each cluster centre is calculated and the respective data point is assigned to the closest cluster. After each observation has been placed in a cluster, the centre of each cluster is recalculated and observations are

reassigned if they are now closer to another cluster centre. This means k-means is sensitive to the order of the data (the order in which each observation is handled), and also to outliers. As the data are not normally distributed and contains outliers these issues may be contributing to this result.

6.3 Summary Of Approaches For Defining the Flourishing Landscape

A number of methods were attempted with a view to satisfactorily summarising the flourishing dataset. This is necessary to facilitate the meaningful visualisation of the countries of the world over time with respect to the dataset.

The methods can be broadly summarised into the traditional principal components analysis approach which extracts new 'summary' dimensions or variables from the variables in the datasets, and a number of clustering methods which attempt to find homogeneous groups, or clusters, within the rows, or in this case countries within the data. These groups can often then be described or profiled by a reduced number of the variables in the dataset. An additional benefit of this second approach is that a small number of clusters may help with the visualisation of a large number of individual data points. This facilitates the implementation of a landscape

approach to measuring human flourishing as first described in the introduction to Chapter 6.

The principal components analysis used the complete set of 107 variables and a number of principal components were extracted. The first six, explaining 65% of the variation in the data, were described in Table 12. The first two were plotted in Figure 8, the *x* and *y*-axes relating largely to variables associated with 'healthcare / education' and 'population / land size' respectively. There is too much unexplained variation in the data to proceed with this approach, however it was a useful comparative technique, the results of which will inform further analysis.

The clustering approaches can be broken into two groups. The spectral clustering approaches – a 2-cluster and 3-cluster solution, and the genetic algorithm approaches – the 'scaled' fitness function, and the 'ranked' fitness function. All four of the clustering approaches, selected a subset of variables to work with. A summary of these subsets can be found in Table 22. The count column indicates the number of times a particular variable appears over the four possible solutions.

Category	Variable Name	Variable Description	Count
Disaster	affected	People affected by natural disasters	1
Economy	BX_KLT_DINV_CD_WD	Foreign direct investment, net inflows (BoP, current US$)	1
	DT_ODA_ALLD_KD	Net official development assistance and official aid received (constant 2008 US$)	2
	FM_AST_DOMS_CN	Net domestic credit (current LCU)	1
	NY_ADJ_DKAP_GN_ZS	Adjusted savings: consumption of fixed capital (% of GNI)	2
	NY_GDP_MKTP_PP_KD	GDP, PPP (constant 2005 international $)	1
	NY_GDP_PCAP_CD	GDP per capita (current US$)	2
	NY_GDP_PCAP_KD	GDP per capita (constant 2000 US$)	1
Education	F15_PRIM_TOT	Percentage of population Female 15+ whose highlest level of education is primary	1
	F15_Prim_Tot	Percentage of population Female 15+ whose highest level of education is primary	2
	F25_NO_SCHOOLING	Percentage of population Female 25+ with no schooling	2
	F25_YEAR_PRIM_SCHOOL	Females 25+ average years of primary schooling	3
	F25_YEAR_TOT_SCHOOL	Females 25+ average years of total schooling	1
	MF15_SEC_TOT	Percentage of population 15+ whose highlest level of education is secondary	1
	MF15_Year_Tert_School	Population 15+ average years of tertiary schooling	1
	MF25_Year_Prim_School	Population 25+ Average years of primary schooling	1
	SE_ENR_PRIM_FM_ZS	Ratio of female to male primary enrollment (%)	1
	SE_PRE_ENRR	School enrollment, preprimary (% gross)	2
	SE_PRM_AGES	Primary school starting age (years)	1
	SE_PRM_DURS	Primary education, duration (years)	1
Environment	NY_ADJ_DCO2_GN_ZS	Adjusted savings: carbon dioxide damage (% of GNI)	1
Health	SH_IMM_IDPT	Immunization, DPT (% of children ages 12-23 months)	1
	SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)	1
	SH_XPD_OOPC_TO_ZS	Out-of-pocket health expenditure (% of total expenditure on health)	1
	SH_XPD_PCAP_PP_KD	Health expenditure per capita, PPP (constant 2005 international $)	2
	SH_XPD_PRIV_ZS	Health expenditure, private (% of GDP)	1
	SH_XPD_PUBL_GX_ZS	Health expenditure, public (% of government expenditure)	3
	SP_DYN_CBRT_IN	Birth rate, crude (per 1,000 people)	1
	SP_DYN_CDRT_IN	Death rate, crude (per 1,000 people)	1
	SP_DYN_LE00_FE_IN	Life expectancy at birth, female (years)	2
Labour	FPR_25N29	Female Labour Force Participation rates Aged 25 to 29	1
	FPR_55N59	Female Labour Force Participation rates Aged 55 to 59	2
	FPR_65P	Female Labour Force Participation rates Aged 65+	1
	MFPR_25N29	Male and Female Labour Force Participation rates Aged 25 to 29	1
	MFPR_30N34	Labour Force Participation rates Aged 30 to 34	1
	MFPR_40N44	Male and Female Labour Force Participation rates Aged 40 to 44	1
	MFPR_45N49	Male and Female Labour Force Participation rates Aged 45 to 49	1
	MFPR_50N54	Male and Female Labour Force Participation rates Aged 50 to 54	1
	MPR_65P	Male Labour Force Participation rates Aged 65+	1
	SL_TLF_CACT_ZS	Labor participation rate, % of total population ages 15+	1
	SL_TLF_TOTL_IN	Labour force, total	2
Land	_eco__agri_area_	Agricultural area square kilometres	1
	_eco__terr_prote	Protected areas - square kilometres	1
	ht_region4	Region: Sub-saharan Africa	2
	lp_lat_abst	Latitude	1
Population	_gen__pop_female	Defacto population as of 1 July of year indicated - female	1
	_gen__pop_rural_	Population residing in rural areas 000's	1
	_gen__pop_total_	Population: de facto population in a country, area or region as of 1 July of the year indicated (000's)	2
	FE_PLURAL	Plurality group: population share of the largest group	2
	MF15_Pop_N_000s	Number of Population 15+	1
	SP_POP_0014_TO_ZS	Population ages 0-14 (% of total)Ê	1
	SP_POP_65UP_TO_ZS	Population ages 65 and above (% of total)	1
	SP_POP_TOTL	Population, total	1
	SP_URB_TOTL_IN_ZS	Urban population (% of total)	1
Production	_res__cereals_ha	Actual cereals harvested sq kilometres	1
	_res__cereals_yi	Actual cereals yielded hectograms per hectare	2
Regime	DPI_SYSTEM2	Regime type (0) Direct presidential (1) Strong president elected by assembly (2) Parliamentary	2
	fh_cl	Civil liberties: 1 (most free) and 7 (least free)	1
	fh_pr1	Political rights: 1 (most free) and 7 (least free).	1
	ht_regtype11	Regime type monarchy	2
	ht_regtype14	Regime type multi-party	2
Religion	LP_PROTMG80	Protestants as % of population in 1980	1
Technology	_gen__mobile_pho	Number of phones per 100 inhabitants	1
	IT_CEL_SETS_P2	Mobile cellular subscriptions (per 100 people)	1

Table 22: The number of times each variable from CD_1995 was selected from the four possible selection methods.

Many variables appear only once, however this is not unexpected. Each solution had a different goal, and the count represents only the variables used to obtain the clusters. Once profiled, variables that may not even appear here may be found to be more effective at describing the differences between the clusters.

The largest focus appears to be on Labour, Health and Education. Within those groups there are variables that allow for gender and age comparisons, potential for equality analysis within those areas.

From here analysis will continue in the form of cluster profiling, evolution and visualisation using the '3-cluster spectral' and the 'Genetic Algorithm ranked fitness function' solutions, with a view to the work by Dasgupta & Ng (2010).

7 Clustering Across Time – Testing And Evaluation Phase

Up to this point, all analysis has been based on data from 1995 (CD_1995). There are a number of possible ways to approach the analysis of data from the years following 1995:

1. Create an individual clustering solution for each year.

2. Score all years following 1995 using the 1995 model. This is an approach commonly used in predictive modelling. A model is built and then future data are scored using the existing model. This requires model performance management such that the model can be amended or rebuilt when performance drops (Chu, Duling, & Thompson, 2007).

3. Build a solution for each year that considers both the current and past data. This could be potentially time consuming, which could be aided if the process is automated.

Option 1 was discarded as it precludes comparison from year to year due to the lack of continuity. Option 2 has the potential for poor snapshot quality, where snapshot quality is a measure of how well C_t, a clustering at time *t,* represents the data at time *t.* For these reasons, only Option 3 was explored. Chakrabarti, Kumar & Tomkins (2006) describe an evolutionary

clustering approach that aims to stay as true to the current data as possible, but should not change too much from one time stamp to the next. They describe the benefits of this approach, including consistency (each solution is similar to the last), noise removal (taking the past into account helps provide protection against outliers in the current data), smoothing (if the clusters shift, the transition should be smooth), and cluster correspondence (current clusters will tend to correspond with past clusters).

Their work used a framework that balances the quality of the current solution (snapshot quality) against the difference between the current clustering solution and the past (history cost) using the formula in equation (3),

$$\sum_{t=1}^{T} sq(C_t, M_t) - cp \sum_{t=2}^{T} hc(C_{t-1}, C_t) \quad \textbf{(3)}$$

where $sq(C_t, M_t)$ is the snapshot quality of the current clustering solution, C_t, with respect to the current data M_t, $hc(C_{t-1}, C_t)$ is the historical cost of the current solution, that is a measure of the distance between C_t and C_{t-1}, and cp is a user defined parameter (cost parameter) that controls the balance between the snapshot quality and historical cost. As cp increases, greater weight is given to matching history, that is, ensuring the current clustering solution is comparable to the one previous. Chakrabarti et al. (2006) applied their framework to two clustering algorithms – agglomerative hierarchical clustering and k-means clustering. At that time

they believed they were the first to undertake such an approach. In the more recent work by (Xu, Kliger, & Hero III, 2011) the authors undertake evolutionary clustering by creating a temporal smoothing parameter using shrinkage estimation (where an estimate is improved by incorporating further information).

There are a number of references, including Naldi, de Carvalho, Campello & Hruschka (2007) that discuss Genetic Algorithms in the context of clustering. However the following analysis has a different approach to using Genetic Algorithms to evolve the clustering approaches described in 6.2 over time.

7.1 Canonical Discriminant Pre-Analysis

In 6.2.1.2.2 an initial canonical discriminant analysis, performed on the 1995 solution, showed some potential. Therefore, investigation continued into canonical discriminant analysis as a way of visualising the cluster analysis results. The original analysis used the model variables - the variables that had been selected for the solution that would provide the best separation between clusters over ALL the variables in the CD_1995 dataset.

To clarify, the fitness of each candidate in the 3-cluster spectral solution was calculated by averaging the results of The Kruskal-Wallis test, Brown-Mood test, Savage test and the Kolmogorov-Smirnov test statistics, for

each of the 107 variables in the CD_1995 dataset. All variables were included to determine the fitness of the candidate solution, but not all variables are necessarily good "separators". Although the variables in the winning solution work in combination to provide the best overall fitness (as just described), they are a limited set. Additionally, although the separation statistics measure a variable's ability to provide good "overall" descriptive separation between clusters, they do not provide information about the separation between all pairs of clusters.

Therefore, an investigation was carried out to find the subset of variables from CD_1995 that provided the best separation based on the 1995 solution described in 6.2.1.2.2, with consideration to the issues described above. Figure 17 is a visual representation, with the left disc showing a representation of the selection of variables in the spectral clustering final solution, (subsequently also used in the canonical discriminant analysis). The disc on the right represents the evaluation of all variables in the dataset, based on information from the spectral clustering final solution, in terms of the ranking of their KS statistics and a visual evaluation of their individual boxplots. The variable numbering has no meaning other than to show an example of category membership.

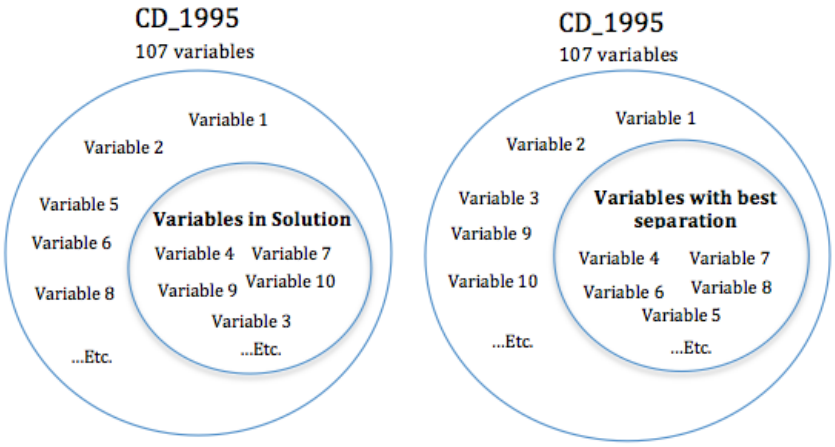


Figure 17: The left disc shows the selection of variables in the final spectral solution vs the right disc which shows the evaluation of all variables in CD_1995 in terms of their ability to provide class separation.

7.1.1 Canonical Discriminant Pre-Analysis Method

As previously mentioned in 6.2.1.1, the KS statistics provide a measure of how well a variable is able to distinguish between clusters over all pairs of clusters in the solution. When more than two categories are present, they do not provide a measure of how well a variable is able to distinguish between the individual pairs of clusters. For example, the spectral clustering final solution has three clusters. A particular variable may have a very low average value in cluster 1, and a very high average value in cluster 3, but cluster 2 may have a very similar average value to cluster 3. In this case the KS statistics may be high due to the wide difference between clusters 1 and 3 and the lack of differentiation between cluster 2 and 3 will be missed.

A boxplot can be helpful in evaluating such cases. For each numeric variable in the dataset, the distribution was plotted by cluster, i.e. the value of that variable for each country in each cluster.

An example can be found in Figure 18 showing boxplots for the percentage of women over 25 whose highest level of education is secondary school.

Canonical Discriminant Analysis was then performed on the variables selected in this way.

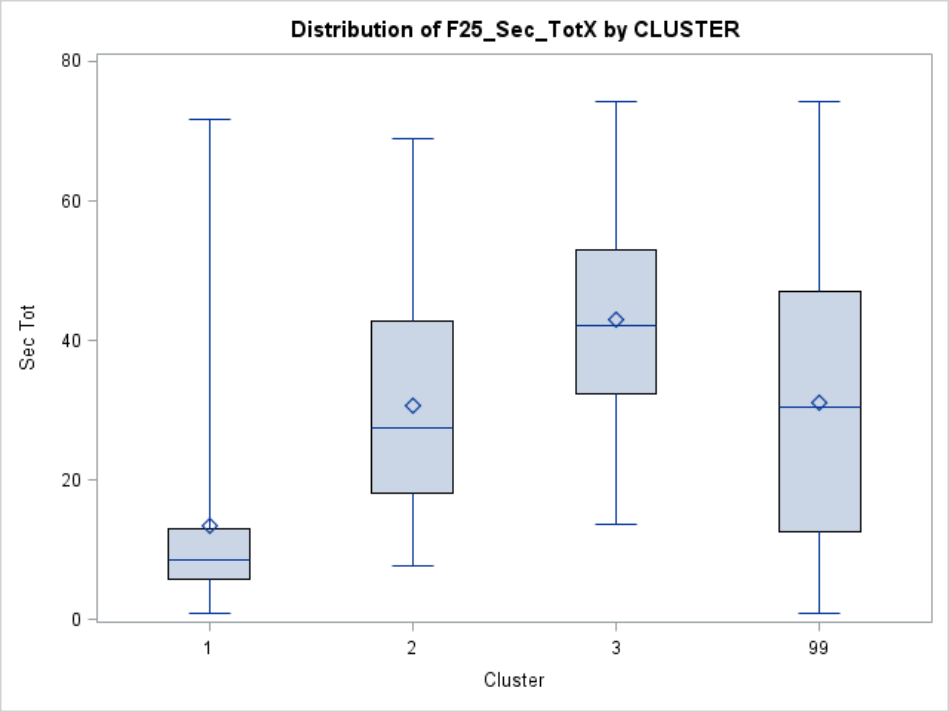


Figure 18: Example boxplots for an individual variable, in this case the number of years at secondary school for women over 25, showing the distribution of the variable by cluster. The value '99' is the distribution over the whole dataset.

7.1.2 Canonical Discriminant Pre-Analysis Results

The separation statistics (the average of The Kruskal-Wallis test, Brown-Mood test, Savage test and the Kolmogorov-Smirnov test statistics) calculated in 6.2.1.1 were examined in terms of their significance. All variables with $p < .001$ for all statistics were kept. The boxplots of the numeric variables were then examined and those that provided good visual separation between all classes were included. For the categorical (dummy) variables, just those significant across all of the separation statistics were included, as boxplot analysis is not appropriate.

Using these criteria, 65 variables were chosen for a canonical discriminant analysis. This number was chosen to ensure a wide range of variables were available in the first algorithm iteration from 1995 to 1996.

Considering the strongest correlations between individual variables and the canonical dimensions, dimension one has a positive correlation with:

- Years of secondary schooling for women over 25
- Years in primary school for women over 25
- Years of total schooling for women over 25
- A regime type of "monarchy"

- Male/female life expectancy
- Male/female years of primary school.

Dimension one has a negative correlation with:

- Men and women in the labour force over 65
- Children as a % of population
- Fertility rate.

To summarise, if all countries in the 3 cluster spectral solution were plotted on these dimensions, moving along dimension one you could expect to see countries with increasing years of schooling and life expectancy, and additionally, countries with decreasing fertility rates, retirees working, children as a proportion of the total population and fertility rate.

Moving along dimension two you would expect to see countries with increasing political freedom and more countries with a monarch. You would expect to see fewer countries with a multi party regime.

7.1.3 Canonical Discriminant Pre-Analysis Summary

Prior to commencing the evolutionary clustering approach described in 7, a further investigation into canonical discriminant analysis as a cluster solution visualisation was undertaken. This was in order to provide a base,

or reference for the variables chosen to participate in the canonical discriminant analyses for the years following 1995.

7.2 Evolutionary Spectral Clustering

The 3-cluster Spectral Clustering algorithm described in 6.2.1.1 was used as the template to this part of the analysis. To summarise, this consisted of a genetic algorithm that I designed and wrote to select the combination of variables from the dataset CD_1995 that would provide the optimal outcome in terms of a spectral clustering algorithm. Optimal was defined as the combination of variables that provided the most differentiation between clusters using the variables in CD_1995. As mentioned in 6.3, the next stage is to amend this process to include years following.

7.2.1 Evolutionary Spectral Clustering Method

The algorithm program used in 6.2.1.1 was copied and renamed with a 1996 extension i.e. starting $t + 1, \text{where } t = 1995$.

Three user-defined parameters were added to control the algorithm progression during crossover and mutation.

1. A maximum number of variables allowed. This is required to limit the variable to observation ratio. At *t*+1, this parameter was assigned the value of 30.

2. A second parameter allowing the user to set a value determining the number of solution variables that must match to those in the previous year's solution. This is to ensure continuity between years and aligns with the snapshot and historical cost trade-off described by Chakrabarti et al. (2006) in equation 3. The variables in the best solution for the current year are compared with the variables in the best solution for the previous year, and if they differ too much, the solution is discarded. At t+1 this parameter was assigned the value of 20.

3. Similarly, a third parameter was added to allow the user to control the number of variables in the canonical analysis required to match the variables in the previous year's canonical analysis. At t+1 this parameter was assigned the value of 30.

Next the method of creating the initial chromosome population was amended so that the probability of a chromosome containing each variable from last year's solution was increased. To clarify, the probability of any variable being included in a chromosome, or tournament candidate, was set at 0.1 to restrict the number of variables in the model to a variable to observation ratio of roughly 0.2. However, if a variable had been in the previous year's (in the first instance 1995) solution, the probability of being selected was set to 0.6. These figures were determined from preliminary testing.

Additionally, the solution chromosome from last year's solution was added to the initial population. In the unlikely event that this chromosome is the current year's winning solution, it would mean there would be high snapshot quality with no historical cost.

The fitness function was amended to include further components, in addition to the separation statistics described in 6.2.1.1, in order to balance the snapshot quality and the historical cost. These included:

1. The number of variables in the current solution.

2. The number of solution variables that match to the previous year's solution.

3. The number of variables selected for canonical analysis that match to the previous year's variables selected for canonical analysis.

These three items, in addition to the separation statistic, were each replaced by the value of their rank in a similar process to that outlined in 6.2.2.1.2. The ideal solution would maximise the match of the current year's solution and canonical variables to the previous year (f_1 and f_2), would also maximise the separation statistics (f_3), and in order to minimise the variable to observation ratio, would minimise the number of variables in the current solution (f_4).

Therefore, the fitness function consisted of accumulating as shown in equation 4.

$$F = f_1 + f_2 + f_3 - f_4 \quad \textbf{(4)}$$

The algorithm continued as in 6.2.1.1, with the fittest chromosome from each tournament being stored until all tournaments have been run. The best of the winners then proceeded to either storage as the 'elite' chromosome, or to crossover and mutation.

An algorithm was created in this way for each year through to 2009. Each algorithm was run several times. Once stable results (by year) were delivered, the results were collated. The collated results included information regarding cluster assignment for each country for each year, a record of the variables used in the solution and a record of the variables to be used in the canonical analysis for each year. The canonical discriminant analysis was performed for each year using this information, and the results plotted.

7.2.2 Evolutionary Spectral Clustering Results

The results of the canonical discriminant analysis for each year were plotted. The results of my analysis can be found at: http://goo.gl/PxSWXG

The visualisation is a Google Charts Motion Chart, an application that is free and open for use by anyone. The list of countries and their associated cluster can also be downloaded from this visualisation.

Figure 19 and Figure 20 are two stills from the visualisation. Over time the visualisation appears to be changing direction. For example, a country that is on the right of the x-axis, (Canonical dimension 1) in 2004, is reflected to the equal but opposite end of the x-axis in 2005.

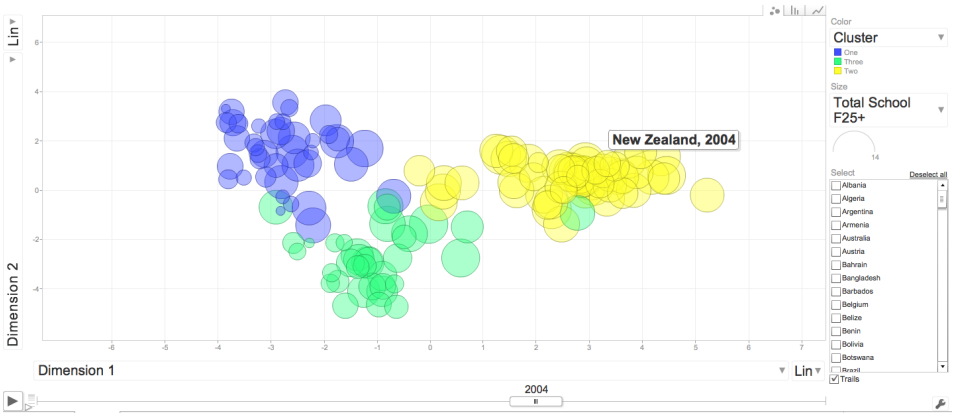


Figure 19: A screenshot from the evolving 3-cluster spectral solution as at 2004.

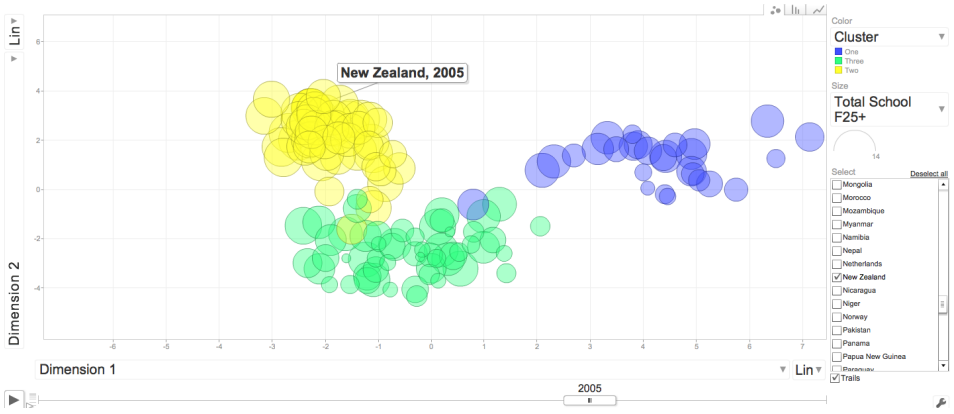


Figure 20: A screenshot from the evolving 3-cluster spectral solution that, when compared to the previous figure, shows the direction reversal from 2004 to 2005.

Investigation of the first two dimensions from the canonical discriminant analysis showed that, although the variables they consisted of were consistent (as required by the algorithm), the direction of the dimension was reversed. For example, in 1999 dimension 1 could be described in part by saying "You would expect to see *increasing* immunization rates", while in 2000 dimension 1 could be described "You would expect to see *decreasing* immunization rates". I concluded that the algorithm assigned the cluster names independently from year to year. That is, cluster 1 in 1999 was not necessarily the companion to cluster 1 in 2000. This meant there was no continuity in the evolution from year to year.

7.2.3 Evolutionary Spectral Clustering Summary

The 3-cluster algorithm from 6.2.1.1, that analysed only CD_1995, i.e. data from 1995 only, was amended to analyse the following years 1996-2009. I modified the algorithm to find a balance between the solution quality for

each current year and continuity with the year previous. A visualisation of the canonical discriminant analysis of the solution showed inconsistencies with cluster naming over time, resulting in an interesting but uninterpretable visualisation.

Spectral clustering is an easily implemented technique, which in this case provided clear clusters. The problem with the cluster continuity could potentially be fixed by ensuring the continuity of cluster names from *t* to *t+1*. However, three clusters were providing a limited categorisation of countries. Therefore, considering the limitations of a 3-cluster solution, and the visualisation difficulty, it was decided to proceed with an evolutionary approach to the genetic algorithm clustering method, ensuring continuity with cluster naming in the future.

8 Clustering Across Time – The Solution

8.1 Evolutionary Clustering – The Genetic Algorithm Approach

The genetic algorithm clustering method described in 6.2.2.1 used a genetic algorithm to find the optimum:

- Combination of variables from the flourishing dataset CD_1995
- Clustering method from a range of seven
- Number of clusters

in order to maximise the winning solution's fitness. Similar to the approach described in 7.2, this was used as a template for the following method.

8.1.1 Method

A canonical discriminant pre-analysis, similar to that described in 7.1, was carried out on the 1995 Genetic Algorithm solution, again to provide a base, or reference, for the variables chosen to be used in the canonical discriminant analyses for the years following 1995.

This time using the Genetic Algorithm from 6.2.2.1, the algorithm was extended similarly to 7.2.1, the "Evolutionary Spectral Clustering Method", including the addition of the extra parameters and the changes to

the initial population. However, this time no adjustments were made to the fitness function.

Considering the results of the previous analysis described in 7.2.2, a control was added to count the cluster matches from the current year to the year previous. For example, if Country A was a member of cluster 2 in 1995 and was a member of cluster 2 in 1996, this was considered a match. It would also have been considered a match if in 1996 Country A was a member of cluster 1 or cluster 3. i.e.

$$C_{t+1} - C_t = \{-1,0,1\} \quad \textbf{(5)}$$

where C_t is a country's cluster assignment at time *t*.

As in 7.2.1 the algorithm was run for each year through to 2009, the canonical discriminant analyses were then undertaken, and the results visualised. My visualisation can be found at:

http://goo.gl/mfKY1m

where it can be seen that adding the parameter to control cluster movement, did not help the visualisation stability. This is because the cluster numbers are not necessarily ordered in a way that means a move as described in equation 5 is a move to a cluster that is a closest neighbour.

A further attempt to stabilise the solution consisted of matching the cluster centroids (using both the mean and median) of the current year's solution to the nearest cluster centroid of the previous year's solution and renaming the current year's cluster accordingly. This also did not help stability.

The conclusion was that the ability of the algorithm to choose the clustering method, in addition to the standardisation approach and outlier treatment for each individual year, was giving a high snapshot quality for the current year, but having too great a historical cost with respect to the previous year.

Therefore, from this point the results of all years from 1995 to 2009 were compared to identify consistencies between the years. The most common clustering method, standardisation approach and outlier treatment were identified, in addition to the 20 most common model variables and the 20 most popular canonical discriminant analysis variables. This was to find a base for all years of genetic algorithm runs. Although this solution could not be guaranteed to remain stable indefinitely, it is reasonable to expect it would for some time, as it is based on a number of year's data.

The data for all years were processed using these static conditions, removing the need for a Genetic Algorithm in the final run. Each cluster median centroid at time *t* was compared to each cluster median centroid at time *t-1* using spearman rank correlation, and each cluster at time *t* was

renamed as per the closest associate at *t-1* if the correlation passed a parameterised threshold. In this case the threshold used was a correlation exceeding 0.95. This meant that more than one cluster in the current year could have the same closest neighbour in the previous year. As such the total number of clusters was dynamic from year to year. Again, using knowledge gained from the analysis so far, the cluster number for each country at time *t* was compared to the cluster number at time *t-1* for the same country. This time a match was defined as a cluster number exactly the same as last year i.e. $C_{t+1} - C_t = 0$. I placed a trigger in my program to stop the run if a parameterised threshold of non-matches was exceeded. In this case the parameter, determined through initial testing, was set such that 70 countries needed to match. This allowed the number of total number of clusters to vary from year to year. After a number of trial runs the maximum number of clusters allowed in the solution was set to five.

8.1.2 Evolutionary Genetic Algorithm Clustering Results

The analysis of the methodology and model variables over the years showed the Flexible Beta clustering method as the most common (with six appearances), and four of those appearances had no standardisation, and no extra outlier suppression.

The Flexible-Beta Clustering method (Lance & Williams, 1967) is an agglomerative hierarchical clustering procedure. That is to say, each observation starts as its own cluster, the two closest clusters are merged,

and the new cluster replaces the two old clusters. There are different ways of calculating the distance between clusters. For the Flexible-Beta Clustering method, the distance is calculated combinatorially (as opposed to directly). The distance is calculated by updating a distance matrix each time two clusters are joined in a way that all measures can be calculated from pre-existing measures. Assuming clusters C_K and C_L merge to form C_M then the distance between C_M and another cluster C_J can be calculated as $D_{JM} = \left(D_{JK} + D_{JL}\right)(1 - b)/2 + D_{KL}b$. It is possible to use the Euclidean distance for this method. In a study examining the effect of various values of the parameter b, (Milligan, 1989) recommends $-0.7 \leq b \leq -0.4$ for data with outliers. As such, a value of b = -0.5 was used in this algorithm.

The top twenty model variables and their number of appearances are shown in Table 23 and the top twenty canonical discriminant analysis variables and their number of appearances is shown in Table 24.

Variable	Appearances	Description
lp_catho80	12	Catholics as % of population in 1980
lp_no_cpm80	11	Other denoms as % of population in 1980:Defined as 100 – lp_catho80 – lp_muslim80 – lp_protmg80.
SP_URB_TOTL_IN_ZS	8	Urban population (% of total)
fe_plural	8	Plurality group: population share of the largest group (ethnic)
SP_DYN_LE00_FE_IN	8	Life expectancy at birth, female (years)
dpi_lipc7	7	Legislative index of political competitiveness: (1) no legislature (2) unelected legislature…(7) largest party got less than 75%
SP_POP_0014_TO_ZS	7	Population ages 0-14 (% of total)Ê
ht_colonial5	7	Colonial origin: (0) Never colonized by a Western overseas colonial power (1) Dutch (2) Spanish(3) Italian (4) US (5) British (6) French (7) Portuguese (8) Belgian (9) British-French (10) Australian
SP_POP_65UP_TO_ZS	7	Population ages 65 and above (% of total)
MF15_Prim_Tot	6	Percentage of population 15+ whose highlest level of education is primary
fh_pr1	6	Political rights: 1 (most free) and 7 (least free).
fe_etfra	6	Ethnic fractionalization: the probability that two randomly selected people from a given country will belong to different ethnic groups
MPR_20N24	6	Male Labour Force Participation rates Aged 20 to 24
MPR_65P	6	Male Labour Force Participation rates Aged 65+
dpi_checks3	5	Number of veto players: increments by one according to competitiveness
MF25_No_Schooling	4	Percentage of population 25+ with no schooling
SH_XPD_PRIV_ZS	4	Health expenditure, private (% of GDP)
fh_cl1	4	Civil liberties: 1 (most free) and 7 (least free)
FPR_25N29	4	Female Labour Force Participation rates Aged 25 to 29
SP_DYN_CBRT_IN	3	Birth rate, crude (per 1,000 people)

Table 23: The most common model variables.

Variable Name	Appearances	Description
_GEN__MOBILE_PHO	12	Number of phones per 100 inhabitants
CHGA_HINST	12	Regime Institutions: (0) Parliamentary democracy (1) Mixed (semi-presidential) democracy (2) Presidential democracy (3) Civilian dictatorship (4) Military dictatorship (5) Royal dictatorship
F15_YEAR_TERT_SCHOOL	12	Females 15+ average years of tertiary schooling
F25_NO_SCHOOLING	12	Percentage of population Female 25+ with no schooling
F25_YEAR_TOT_SCHOOL	12	Females 25+ average years of total schooling
FH_CL	14	Civil liberties: 1 (most free) and 7 (least free)
FH_PR	13	Political rights: 1 (most free) and 7 (least free).
HT_COLONIAL	13	Colonial origin: (0) Never colonized by a Western overseas colonial power (1) Dutch (2) Spanish(3) Italian (4) US (5) British (6) French (7) Portuguese (8) Belgian (9) British-French (10) Australian
LP_CATHO80	13	Catholics as % of population in 1980
LP_MUSLIM80	12	Muslims as % of population in 1980
MF25_NO_SCHOOLING	12	Percentage of population 25+ with no schooling
MPR_65P	13	Male Labour Force Participation rates Aged 65+
NY_GDP_PCAP_KD	11	GDP per capita (constant 2000 US$)
SH_XPD_PCAP_PP_KD	12	Health expenditure per capita, PPP (constant 2005 international $)
SP_DYN_CBRT_IN	14	Birth rate, crude (per 1,000 people)
SP_DYN_LE00_FE_IN	13	Life expectancy at birth, female (years)
SP_DYN_LE00_MA_IN	12	Life expectancy at birth, male (years)
SP_POP_0014_TO_ZS	14	Population ages 0-14 (% of total)Ê
SP_POP_65UP_TO_ZS	14	Population ages 65 and above (% of total)
WDI_FR	13	Fertility rate: births per woman Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with current age-specific fertility rates.

Table 24: The most common canonical discriminant analysis variables.

The run produced the output shown in Table 25 where the evolution of the cluster universe can be seen from year to year. For 1995 and 1996 there are five clusters, in 1997 this reduces to three, indicating that two clusters were reassigned based on their correlation with clusters in the previous year. In

1998 the cluster 19981 and 19982 appear, the prefix indicating the year they occurred. Of these two additions, only 19982 remains in 1999 and both disappear in 2000. In 2002 a new cluster appears, and then in 2005 a further addition. These two remain until completion of the algorithm.

Number of Countries in each Cluster by Year

		Cluster								
		1	**2**	**3**	**4**	**5**	**19981**	**19982**	**20021**	**20051**
Year	**1995**	22	31	16	35	23				
	1996	36	17	16	36	23				
	1997	34			60	34				
	1998	34			37	20	18	19		
	1999	41			51	17		19		
	2000	41			51	36				
	2001	39			51	38				
	2002	54			37	18			19	
	2003	42			50	16			20	
	2004	38			53	17			20	
	2005	38			35	17			20	18
	2006	36			35	28			11	18
	2007	39			36	24			11	18
	2008	39			36	24			11	18
	2009	39			35	24			11	19

Table 25: The distribution of country counts across each cluster by year.

Further analysis revealed a high correlation between clusters 2, 19981 and 20051, and likewise between clusters 3, 19982 and 20021. These were not renamed in the process because the algorithm only looked one year back, however this produced some interesting findings summarised next.

8.1.3 Evolutionary Genetic Algorithm Clustering Summary

The algorithm from 6.2.2.1 was extended to run for all years with an added parameter to control how many countries could move between clusters

from year to year. This approach did not produce stable results. Next, clusters were renamed based on their closest historical neighbour but this approach also did not produce stable results. The final solution was arrived at using the most common clustering methodology, chosen over all years by the Genetic Algorithm. The cluster landscape evolved, and countries moved within this space from year to year. Clusters were renamed to their closest historical match in order to ensure continuity. This is in contrast to the more traditional approach of "scoring" future data based on an original solution, which also provides historical continuity but has the potential to deliver poor snapshot quality.

8.2 Profiling

Given the small amount of data, solution validation using a training and testing set was not undertaken. However, noting the stability of each country's cluster membership from year to year does provide some validation. A further way to validate the solution is to profile the clusters using the internal data (the data that were used to create the solution, in this case the flourishing dataset), and external data (further data that can be matched to cluster members which may be of interest in describing the clusters), noting if the qualities of a cluster description make sense in the context they appear.

8.2.1 Profiling Method

Firstly, the 'internal' cluster descriptions were produced. Where appropriate, boxplots similar to Figure 18 were produced profiling the clusters against each of the variables that had been chosen for the canonical discriminant analysis. The example in Figure 21 shows the distribution of 'Years of education for women over 25+' for each cluster. Comparison of the distribution of countries in clusters 2, 19981 and 20051 is particularly interesting in that they are similar, but look to be slightly increasing over time as clusters 19981 and 20051 appear in 1998 and 2005 respectively. Frequency tables were also produced to examine the distribution spread of certain categorical variables. The boxplots and frequency tables used can be found in Appendix 5.

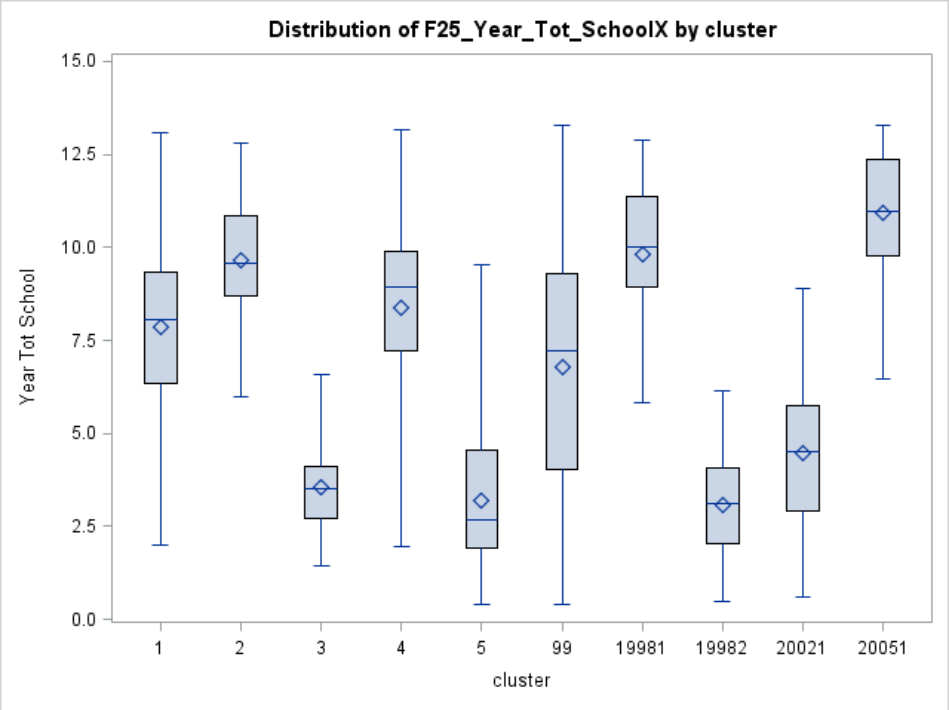


Figure 21 Distribution of years of total schooling for women over 25

8.2.1.1 The World Values Survey

The first external dataset combined with the clustering solution was the World Values Survey (WVS), first mentioned in 4.3. Up to this point, the data being analysed are obtained at country level. The data from the World Values survey are obtained at individual person level and as such need pre-processing before being matched to the respective countries, their clusters and associated internal data.

Using information from the WVS website (Medrano, 2005) I weighted and summarised the data at the country and year level to obtain the mean and median, initially for the life satisfaction, feelings of happiness and state of

health variables. To clarify, respondents are asked to give a rating in response to the following questions:

1. Life Satisfaction: All things considered, how satisfied are you with your life as a whole these days? 1 dissatisfied to 10 satisfied.

2. Feelings of Happiness: Taking all things together, would you say you are: 1 Very Happy to 4 Very Unhappy.

3. State of Health: All in all, how would you describe your state of health these days? Would you say it is...1 Very Good to 5 Very Poor.

The WVS takes place every five years, (World Values Survey, 2008), however individual countries are interviewed in different years within each five year interval such that there are data available for each year. Each occurrence is termed a 'Wave'. I validated the assumption that each country's respondents were interviewed in only one year per wave and then obtained the mean and median for each country / year combination. Up to this point there have (generally) been data for each country for each year. However this is not the case for the WVS. Given analysis is now being done at cluster rather than country level, this is less of a problem and although not ideal, it is unavoidable.

Each country, year combination of WVS data was matched with each country, year, cluster and respective internal variables combination. The incremental change in GDP and Healthcare spend was calculated for each country over the years data were available. Tables of the Spearman Rank and Pearson correlations were produced. These were done by cluster and treated the life satisfaction, happiness and state of health variables separately. Correlations significant at $p < 0.05$ level (in both the Pearson and Spearman correlation tables) were used in the cluster descriptions. Figure 22 shows an example of the Spearman table for cluster 1 and the Life Satisfaction variable. The 'mean' and 'pctl_50' variables are the mean and median Life Satisfaction respectively.

Spearman Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Mean	Pctl_50	NY_GDP_PCAP_KD	SH_XPD_PCAP_PP_KD	rel_inc_health	rel_inc_GDP	EN_ATM_CO2E_PC	NY_ADJ_DCO2_GN_ZS	NY_ADJ_DKAP_GN_ZS	NY_ADJ_DNGY_GN_ZS
Mean	1.00000	0.97059	0.28698	0.16176	0.11176	-0.18690	0.10357	-0.05588	-0.03297	0.58529
		<.0001	0.2812	0.5495	0.6803	0.4882	0.7134	0.8371	0.9109	0.0172
	16	16	16	16	16	16	15	16	14	16
Pctl_50	0.97059	1.00000	0.25901	0.13824	0.19412	-0.15305	0.13571	0.03235	-0.04176	0.62353
50% Percentile	<.0001		0.3327	0.6097	0.4713	0.5715	0.6296	0.9053	0.8873	0.0099
	16	16	16	16	16	16	15	16	14	16
NY_GDP_PCAP_KD	0.28698	0.25901	1.00000	0.80206	0.35026	0.54639	0.74173	-0.28109	0.65714	-0.07211
GDP per capita (constant 2000 US$)	0.2812	0.3327		0.0002	0.1835	0.0285	0.0015	0.2916	0.0107	0.7907
	16	16	16	16	16	16	15	16	14	16
SH_XPD_PCAP_PP_KD	0.16176	0.13824	0.80206	1.00000	0.31471	0.44297	0.73929	-0.36471	0.76703	-0.36176
Health expenditure per capita, PPP (constant 2005 international $)	0.5495	0.6097	0.0002		0.2352	0.0857	0.0016	0.1649	0.0014	0.1686
	16	16	16	16	16	16	15	16	14	16
rel_inc_health	0.11176	0.19412	0.35026	0.31471	1.00000	0.55040	0.64286	-0.22353	0.58242	0.11471
	0.6803	0.4713	0.1835	0.2352		0.0272	0.0097	0.4053	0.0289	0.6723
	16	16	16	16	16	16	15	16	14	16
rel_inc_GDP	-0.18690	-0.15305	0.54639	0.44297	0.55040	1.00000	0.57015	-0.18690	0.41978	-0.15453
	0.4882	0.5715	0.0285	0.0857	0.0272		0.0265	0.4882	0.1351	0.5677
	16	16	16	16	16	16	15	16	14	16
EN_ATM_CO2E_PC	0.10357	0.13571	0.74173	0.73929	0.64286	0.57015	1.00000	-0.02143	0.74176	-0.01786
CO2 emissions (metric tons per capita)	0.7134	0.6296	0.0015	0.0016	0.0097	0.0265		0.9396	0.0037	0.9496
	15	15	15	15	15	15	15	15	13	15
NY_ADJ_DCO2_GN_ZS	-0.05588	0.03235	-0.28109	-0.36471	-0.22353	-0.18690	-0.02143	1.00000	-0.38901	0.40588
Adjusted savings: carbon dioxide damage (% of GNI)	0.8371	0.9053	0.2916	0.1649	0.4053	0.4882	0.9396		0.1692	0.1188
	16	16	16	16	16	16	15	16	14	16
NY_ADJ_DKAP_GN_ZS	-0.03297	-0.04176	0.65714	0.76703	0.58242	0.41978	0.74176	-0.38901	1.00000	-0.55165
Adjusted savings: consumption of fixed capital (% of GNI)	0.9109	0.8873	0.0107	0.0014	0.0289	0.1351	0.0037	0.1692		0.0408
	14	14	14	14	14	14	13	14	14	14
NY_ADJ_DNGY_GN_ZS	0.58529	0.62353	-0.07211	-0.36176	0.11471	-0.15453	-0.01786	0.40588	-0.55165	1.00000
Adjusted savings: energy depletion (% of GNI)	0.0172	0.0099	0.7907	0.1686	0.6723	0.5677	0.9496	0.1188	0.0408	
	16	16	16	16	16	16	15	16	14	16

Figure 22: An example correlation table, in this case for cluster 1, showing the correlation of the mean and median life satisfaction (mean, pctl_50) with other variables.

The WVS data were examined for completeness, in order to determine which variables to use in the profiling in addition to life satisfaction, happiness and health. There are missing data in the WVS dataset where, for example, countries and their citizens have not been asked certain questions, or questions may not have been asked for all waves. Table 26 shows short descriptions of the variables examined.

Variable Label
Satisfaction with your life
How much freedom of choice and control
Membership consumer organisation
Self positioning in political scale
Income equality
Current society: Egalitarian vs. competitive society
Current society: Extensive welfare vs. low taxes
Current society: Regulated vs. responsible society
Society aimed: egalitarian vs. competitive
Society aimed: extensive welfare vs. low taxes
Society aimed: regulated vs. responsible society
Thinking about meaning and purpose of life
I see myself as a world citizen
I see myself as member of my local community
I see myself as citizen of the [country] nation
I see myself as an autonomous individual
SURVIVAL/SELF-EXPRESSION VALUES
TRADITIONAL/SECULAR RATIONAL VALUES
Nature of tasks: manual vs. Cognitive
Nature of tasks: routine vs. Creative
Nature of tasks: independence
Post-Materialist index 12-item
Post-Materialist index 4-item

Table 26 :The variables that were examined from the World Values Survey.

Examination of these variables was done using plots over time of each individual variable's mean and median by cluster to see where each cluster ranked within a variable relative to each other cluster. Appendix 5

contains the plots relating to the variables used in the cluster descriptions. Figure 23 is an example of the plot of the mean score by cluster of the survival/self expression dimension of the Global Cultural Map. Inglehart & Welzel (2010) used Factor Analysis to derive two dimensions from twelve 'attributes' in the WVS. These are a traditional / secular dimension and a survival / self-expression dimension. A low score indicates the survival end of the spectrum while a high score indicates the self-expression end of the dimension. The WVS data were not used in the creation of the clusters, however there is separation between several of the cluster's mean scores. Missing data points indicate no data available for that cluster for that year.

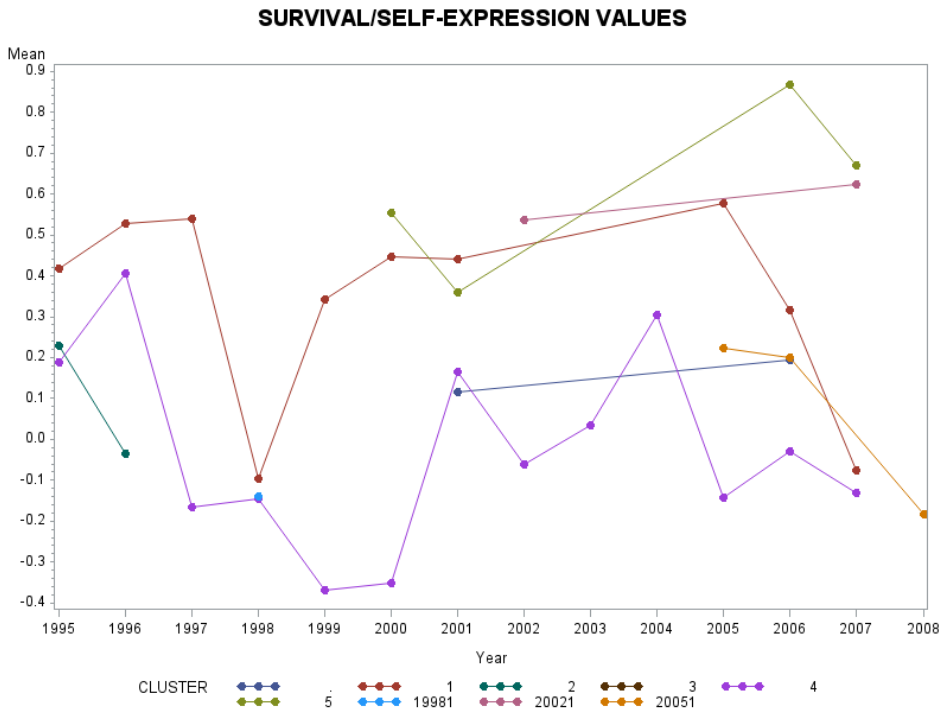


Figure 23: The mean survival/self expression score plotted by cluster and year.

8.2.1.2 The Happy Planet Index

The cluster, country dataset for the year 2009 was combined with the data from the Happy Planet Index (HPI), first mentioned in 3.3. This dataset also contains the core information that constitutes the HPI including a measure of life satisfaction / wellbeing from The Ladder of Life question on the Gallup World Poll, and global footprint (Abdallah et al., 2012). For simplicity in output interpretation, clusters 20051 and 20021 were renamed clusters 2 and 3 respectively. Because only one year's data were available, the values associated with countries within each cluster were ranked to facilitate visualisation of the clusters.

8.2.1.3 University Of Texas Inequality Project

The Estimated Household Income Inequality data (EHII) from the University of Texas Inequality Project (Galbraith, 2010) was matched to the country, cluster year information from 1995 onwards, and the Happy Planet Index indicators for 2009 only. The Income Inequality index is available from 1963 onwards and uses a measure derived from Theil's T Statistic. The Spearman and Pearson correlations, and the plots of inequality by cluster, and with respect to NZ, produced findings shown in 8.2.2.3.

8.2.2 Profiling Results

8.2.2.1 WVS

Not all variables analysed from the WVS provided visual separation between the clusters. Those that did, such as Figure 23, can be found in Appendix 5 and were included in the cluster descriptions to follow.

8.2.2.2 Happy Planet Index

Analysis of the Happy Planet Index (2009 only) produced mixed results. Figure 24 shows the plot of each cluster's Happy Planet Index distribution where each point represents a country and its associated rank within the cluster. There is overlap between the clusters, however when ranked, there is a noticeable difference between cluster 3 (high) and cluster 5 (low). The distributions of the remaining clusters appear quite similar. Figure 25 and Figure 26 show the HPI broken into two components – Global Footprint and Life Satisfaction respectively. Countries in cluster two tend to have the largest global footprint, wealth and additionally life satisfaction, but the 'gap' between that of cluster two's global footprint and the other clusters' is greater than the same 'gap' for life satisfaction. This is interesting in the context of (Myers, 2000), research at individual level showing a decrease in the growth of life satisfaction past a certain income.

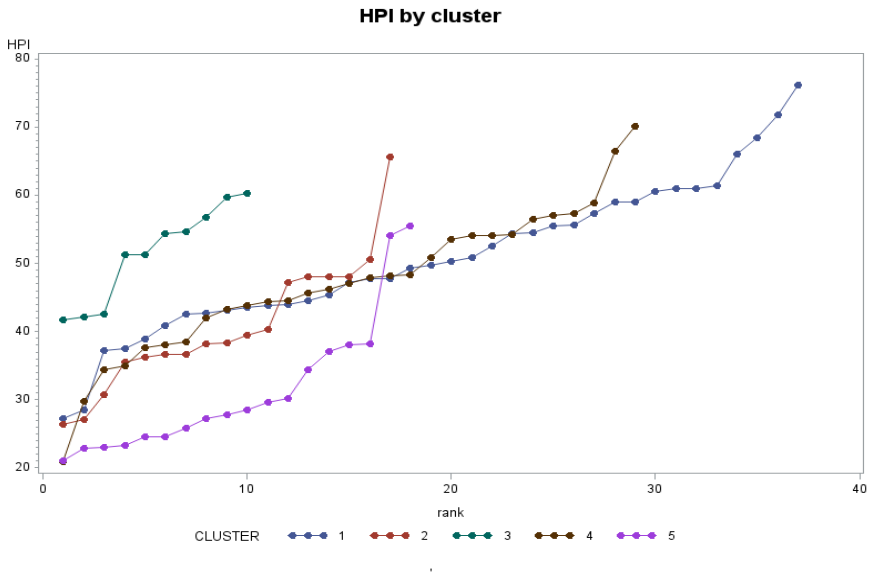


Figure 24: A plot of Happy Planet Index by cluster, where each point represents a country and its associated Happy Planet Index (actual, and then ranked within its cluster).

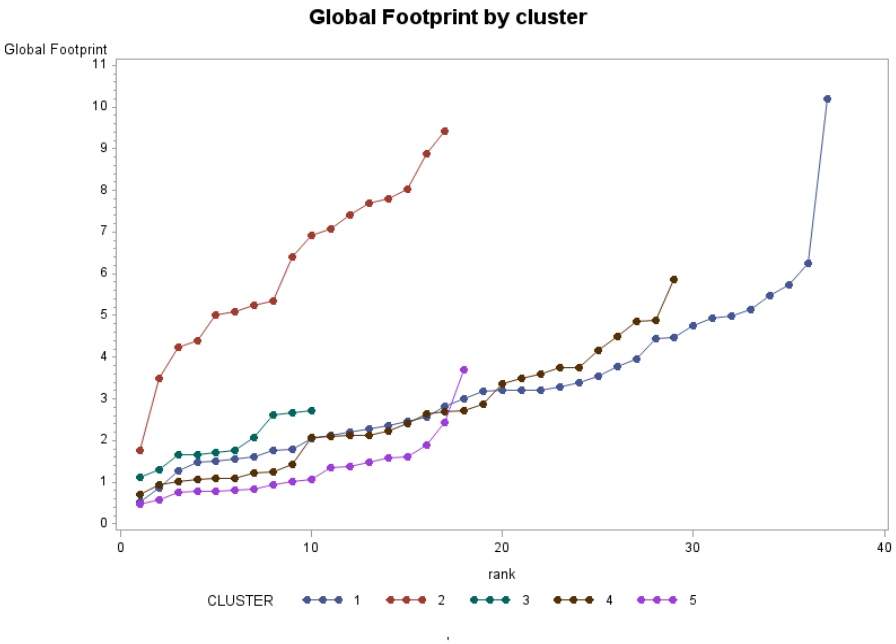


Figure 25: A plot of Global Footprint by cluster, where each point represenst a country and its associated Global Footprint (actual, and then ranked within its cluster).

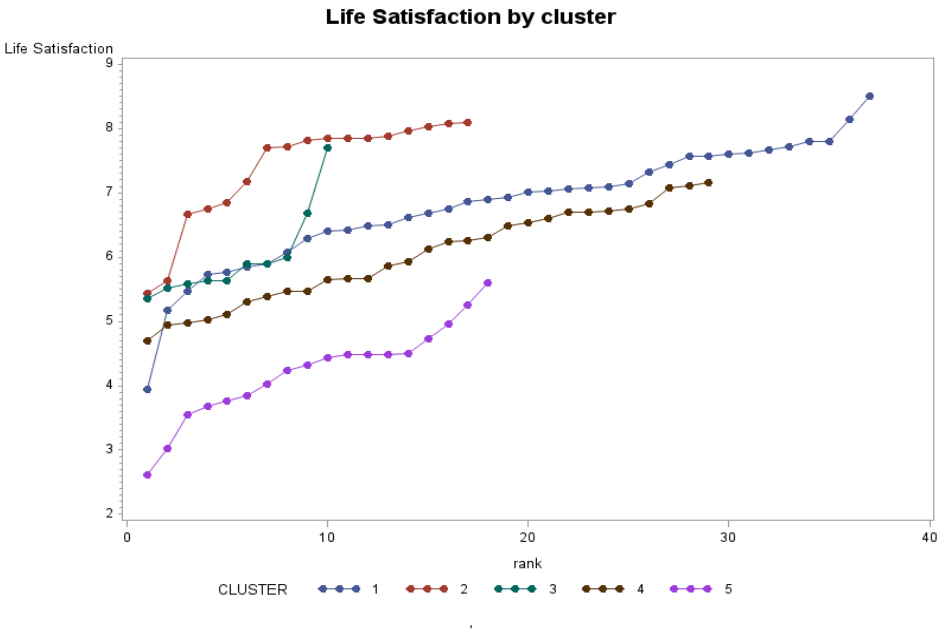


Figure 26: A plot of Life Satisfaction by cluster, where each point represents a country and its associated life satisfaction (actual and then ranked within each cluster).

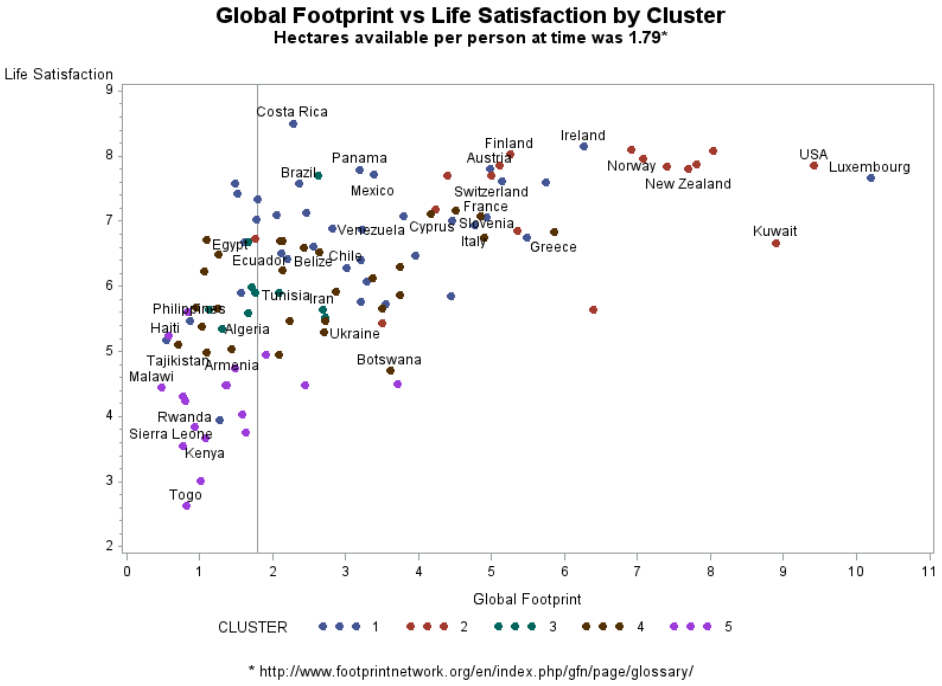


Figure 27: A plot of Global Footprint vs Life Satisfaction for each country by cluster.

An interesting question to ask is, is there a point for an individual country, at which consuming more does not substantially improve life satisfaction? Figure 27 shows the relationship between life satisfaction and global footprint, by cluster. Although there is no clear separation between clusters, again there appears to be a 'ceiling' of life satisfaction and a decrease in the growth of life satisfaction past a global footprint of around 3. Countries with the highest life satisfaction seem to be mostly in cluster 1 or cluster 2, however those in cluster 1 seem to generally have a lower global footprint. Cluster 5 has the lowest life satisfaction and global footprint overall. The vertical line indicates the number of global hectares available at the time the data were produced. The difference between the line and countries to the right of that line indicates how many more global hectares per person a country is using, than is actually available.

8.2.2.3 University Of Texas Inequality Project

Figure 28 shows the boxplots of the cluster distributions in terms of inequality for the years 1995 - 2009. Although there is some overlap between clusters, cluster 5 in general has higher inequality than all clusters, only overlapping cluster 3 (the next highest inequality in general). Cluster 2 has the lowest inequality overall.

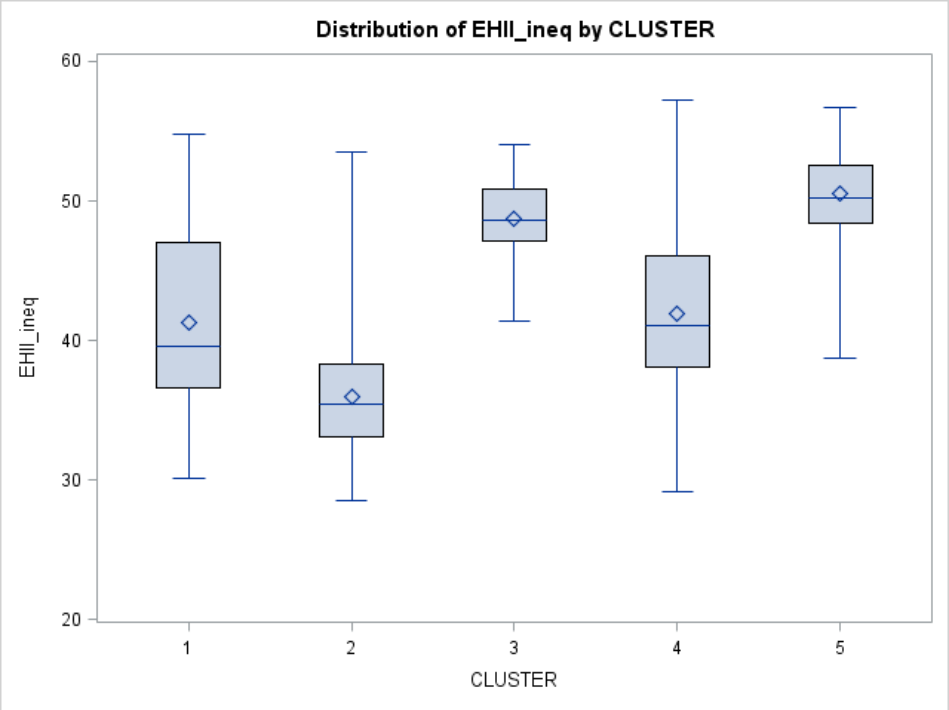


Figure 28: Boxplots showing the distribution of inequality by cluster.

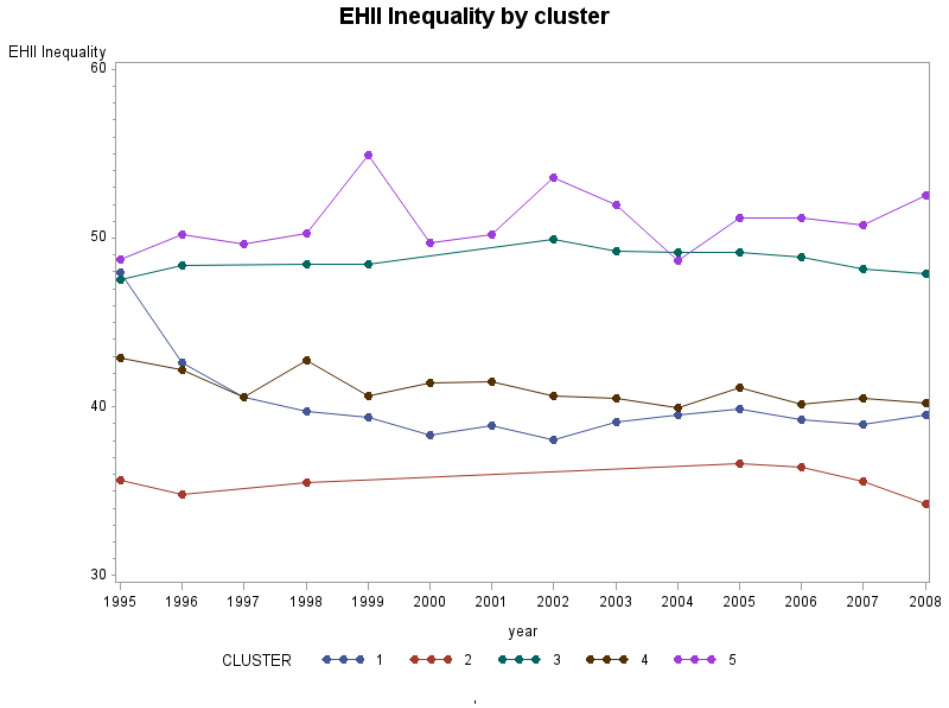


Figure 29: A plot of the median inequality value by cluster.

Figure 29 shows the inequality medians of each cluster over time. There are missing data points where there are gaps in the inequality data.

Table 27 shows the list of correlations by cluster between the listed variable and inequality which are significant at a level of p=0.05 or stronger for both the Spearman and Pearson correlations.

When all clusters are combined there is a significant negative correlation between inequality and life satisfaction. That is, as inequality rises, life satisfaction will tend to decrease, as will life expectancy, happy life years, GDP and the human development index.

When the analysis is performed by cluster, there are insufficient data and / or insignificant correlations for clusters 2, 3 and 5. However, Figure 28 suggests less of a spread of inequality within these clusters i.e. they are generally high or generally low.

NZ is a member of several clusters over the years 1995 to 2008. In particular Cluster 2 from 1995-1998, 2005-2007, Cluster 1 in 2002 and Cluster 4 from 2003-2004.

All Clusters	R	p*
Pearson Correlation Coefficients, N = 91		
Prob > |r| under H0: Rho=0		
Life Sat (0-10)	-0.50738	<.0001
Life Exp (years)	-0.56517	<.0001
Footprint (g ha /cap)	-0.57527	<.0001
Happy Life Years	-0.56927	<.0001
GDP per capita ($ PPP)	-0.66333	<.0001
Human Development Index	-0.69521	<.0001
Spearman Correlation Coefficients, N = 91		
Life Sat (0-10)	-0.51586	<.0001
Life Exp (years)	-0.62349	<.0001
Footprint (g ha /cap)	-0.65627	<.0001
Happy Life Years	-0.56821	<.0001
GDP per capita ($ PPP)	-0.71153	<.0001
Human Development Index	-0.73682	<.0001
CLUSTER=1		
Pearson Correlation Coefficients, N = 33		
Life Exp (years)	-0.72252	<.0001
Footprint (g ha /cap)	-0.68252	<.0001
Happy Life Years	-0.45859	0.0073
Happy Planet Index	0.56999	0.0005
HPI rank	-0.62566	<.0001
GDP per capita ($ PPP)	-0.75494	<.0001
Human Development Index	-0.81952	<.0001
Spearman Correlation Coefficients, N = 33		
Life Exp (years)	-0.71051	<.0001
Footprint (g ha /cap)	-0.79311	<.0001
Happy Life Years	-0.46591	0.0063
Happy Planet Index	0.61664	<.0001
Happy Planet Index Ranked	-0.61664	0.0001
GDP per capita ($ PPP)	-0.83456	<.0001
Human Development Index	-0.80749	<.0001
CLUSTER=2 None Significant		
CLUSTER=3 Insufficient Data		
CLUSTER=4		
Pearson Correlation Coefficients, N = 30		
Life Sat (0-10)	-0.43161	0.0172
Life Exp (years)	-0.3887	0.0338
Happy Life Years	-0.44435	0.0139
Human Development Index	-0.47614	0.0078
Spearman Correlation Coefficients, N = 30		
Life Sat (0-10)	-0.38954	0.0334
Life Exp (years)	-0.37335	0.0421
Happy Life Years	-0.42113	0.0205
Human Development Index	-0.43626	0.0159
CLUSTER=5 Insufficient Data		

*Prob > |r| under H0: Rho=0

Table 27: Correlations between inequality and various other variables showing only those significant at p<=0.05.

Figure 30 shows NZ's inequality from 1995-2008 compared with the median of each of these clusters. NZ's inequality is always higher than the median of cluster 2 and lower than the medians of clusters 1 and 4, until 2005 when it becomes, and continues to be higher, than all three, cluster medians.

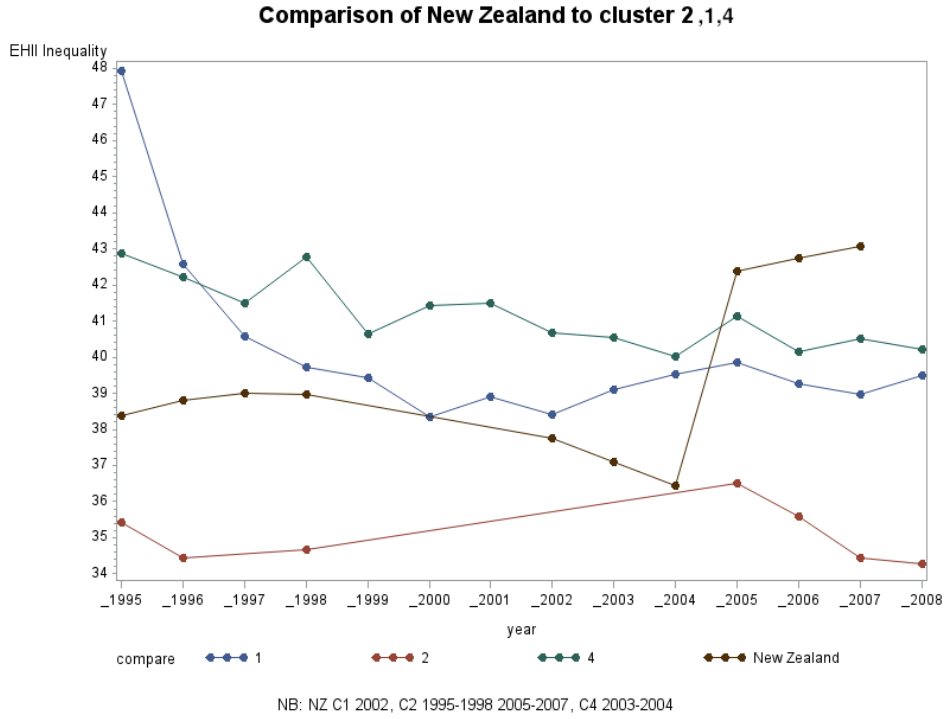


Figure 30: Inequality in New Zealand by year, as it relates to the median inequality of the clusters NZ is, at various times, a member of.

The inequality data do have some missing data points but on the other hand are interesting because it contains data from the 1960s. Figure 31 shows a plot of NZ's inequality over this period compared to Australia and the UK. It is unfortunate that data for Australia are unavailable after 2001 as it would be interesting to compare with the increase in NZ's inequality from

2004 to 2005 following a dip from 2002-2004. It can be seen there is no similar increase in the UK data. I am unsure as to the reason behind the increase. However, it is important to note that the increase follows a dip and appears more in line with the trend prior to the prior data point in 1998.

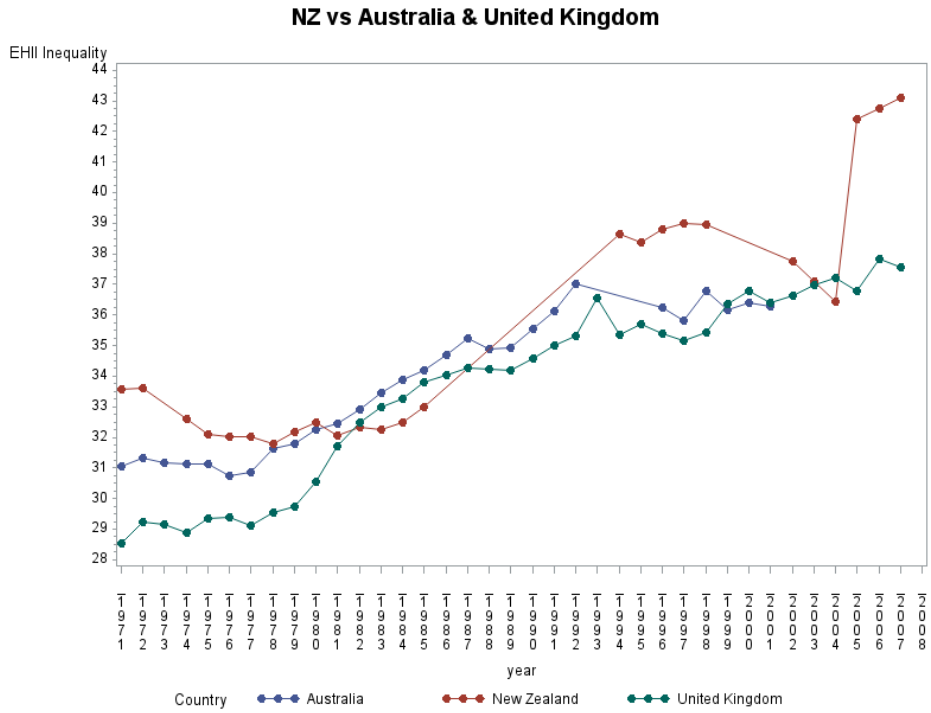


Figure 31: Inequality in New Zealand compared, by year, with inequality in Australia and the United Kingdom.

8.2.2.4 The Clusters

The cluster descriptions, whilst changing slightly over time, were relatively stable. As depicted in Table 25, sometimes clusters were absorbed into other clusters as time progressed and at times new clusters were created - with strong associations to previous clusters.

Figure 32 is The World Values Survey Cultural map 2005-2008 (Inglehart & Welzel, 2010). The data are all of the countries that took part in Wave 5 of the WVS, plotted by the two dimensions first mentioned in reference to Figure 23.

This map is mentioned in the cluster descriptions that follow, and information supporting the descriptions can be found in Appendix 5.

8.2.2.4.1 Cluster One – Traditional Forward Thinkers

In general, and relative to the other clusters, women over 25, who live in countries belonging to this cluster, will have less total schooling than the world median, there will relatively and generally be a very high proportion of Catholics, a very low proportion of Muslims and civil liberties will tend to be relatively low. Countries in this cluster will most likely have never been colonised, or may have been colonised from Spanish origin.

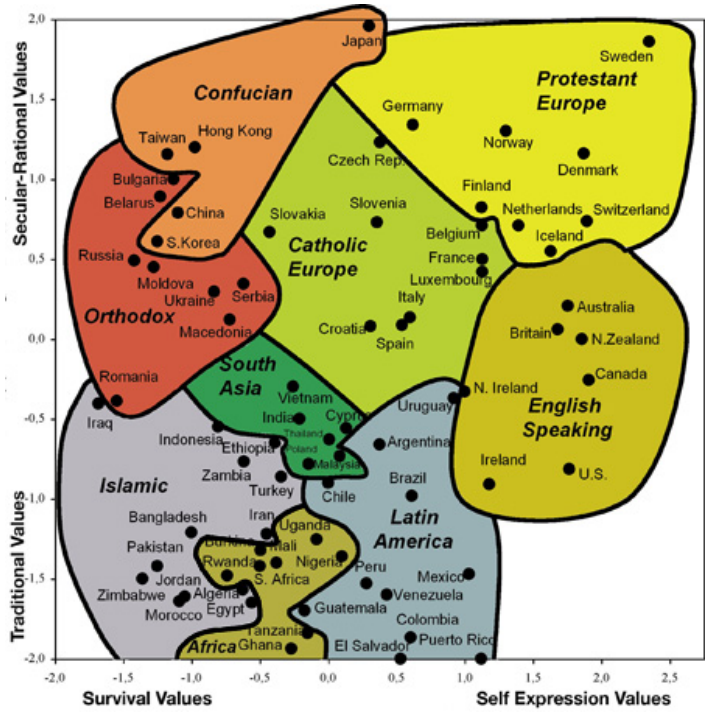


Figure 32: The World Values Survey Cultural Map 2005-2008.

This cluster seems to have a relatively reasonable satisfaction of life, and also reasonably high feelings of personal choice and control - at odds with the actual civil liberties of this cluster. This cluster tends to be politically central and tending towards the left in more recent years. Members of countries in this cluster tend to rate their country as egalitarian (the gap between rich and poor is small), and have extensive welfare rather than low taxes as their ideal view. People in this cluster will tend to believe that society should continue to be aiming to a more egalitarian view. Of all clusters, this cluster tended to be least likely to agree with the sentiment "I see myself as an autonomous individual". This cluster is tending towards the self-expression end of the survival/self expression values dimension,

but is not the cluster furthest along, and is slightly towards the traditional end of the traditional / secular axis. This cluster thinks a lot about the meaning of life.

In respect to the WVS Cultural map shown in Figure 32, countries in this cluster will tend to be in the lower middle area, described on the map as Latin America.

8.2.2.4.2 Cluster Two - Satisfied, Free And Central

In general, and relative to the other clusters, women who live in countries belonging to this cluster, will have more total schooling and more years of tertiary schooling than the world median, less men over 65 in employment, higher GDP and healthcare spending per capita, a lower birth-rate and a longer life expectancy. The ratio of under 14s to over 65s will be small, the number of phones per 100 inhabits will be low, the proportion of Muslims in the population will be low, and civil liberties and political rights will be high. They will likely have never been colonised.

This cluster appears to have the highest satisfaction with life and feels they have a high amount of freedom and control. Their views tend towards the political centre, and around the middle of the income inequality / bigger income differences scale. People in this cluster tend to agree strongly with the statement "I see myself as an autonomous individual".

In respect to the WVS Cultural Map countries in this cluster tend to be in the Protestant Europe / English speaking area.

8.2.2.4.2.1 Cluster 19981

Cluster 19981 can be described very similarly to cluster two, but the description could be said to be even stronger. This is an example of the clusters are changing over time but still remaining essentially the same.

8.2.2.4.2.2 Cluster 20051

Cluster 20051 is very similar to Cluster 19981 and Cluster two, but even stronger. People in this cluster appear the most satisfied with life and feel they have a high degree of freedom and control. They tend slightly to the left politically and prefer more equality rather than larger differences. This cluster appears to do the most cognitive and creative work. Although still materialist, this cluster appears to be the least materialist of all.

8.2.2.4.3 Cluster Three - Religious Traditionals

In general, and relative to the other clusters, women who live in countries belonging to this cluster will have less total and tertiary schooling than the world median, there will tend to be less GDP and healthcare spend per capita and a higher birth-rate. There will tend to be a large proportion of under-14s to over-65s, a low number of phones per 100 inhabitants, a very low proportion of Catholics and a very high proportion of Muslims. There will tend to be a fairly high fertility rate and the lowest of all civil liberties

and political rights. If colonised, countries in this cluster will most likely have French or British colonial origin.

People in this cluster tend to have relatively low life satisfaction and perceived freedom and control. They tend to the middle of the political spectrum and to think a lot about the meaning and the purpose of life. On other variables already mentioned in previous clusters, and where data are available, they seem to remain around the middle.

In respect of the WVS Cultural Map, countries in this cluster tend to be in the lower left.

8.2.2.4.3.1 Cluster 19982

Cluster 19982 is very similar in description to cluster 3 where women who live in countries belonging to this cluster will have less schooling than the world median, there will tend to be less GDP and healthcare spend per capita and a higher birth-rate. There will tend to be, relative to other clusters, very low civil liberties, political rights and life satisfaction.

8.2.2.4.3.2 Cluster 20021

Cluster 20021 can be described very similarly to Cluster 19982 and Cluster 3, but with perhaps a slight improvement in conditions. There appears to be a substantial amount of manual, routine work relative to other clusters. People in this cluster will tend to prefer larger pay differences as incentives (as opposed to equality).

8.2.2.4.4 Cluster Four – Middling And Competitive

In general, and relative to the other clusters, women who live in countries belonging to this cluster will have more total schooling than the world median. There will tend to be a smaller proportion of men over 65 in employment, and there is a relatively low fertility rate. These countries are most likely to have never been colonised or have been colonised by the British.

People in this cluster are 'middling' regarding satisfaction of life and feelings of freedom and control. They tend to a central or slightly right political view and believe their society is currently more competitive than egalitarian, and has low taxes rather than a welfare state. They feel society should be aiming to be more competitive.

In respect of the WVS Cultural Map, countries in the cluster will tend to be in the upper left, ex-communist area.

8.2.2.4.5 Cluster Five - Struggling Traditionals

In general, and relative to the other clusters, women who live in countries belong to this cluster will have the least total and tertiary schooling and the most men over 65 in employment. Countries in this cluster will tend to have the lowest GDP and healthcare spending per capita and the lowest life expectancy for both men and women. They will tend to have a very high birth rate and the ratio of under-14s to over-65s will be the highest. If

colonised, countries in this cluster will most likely have French, British or Belgian colonial origin.

Members of this cluster are the least satisfied with their lives and feel they have the least freedom and control. They think about the meaning and purpose of life the most. They agree strongly with the statement that they are a world citizen, but also feel strongly that they are autonomous individuals. They are the most materialist on the post-materialist index and do the most routine / manual work. Countries in Cluster 5 tend to be in the far lower left of the WVS Cultural map.

8.2.3 Profiling Summary

A number of external sources of data including variables from the WVS, HPI and EHII were used to produce profiles of the clusters created in 8.1. Additionally, the clusters were profiled using their own internal data.

Five clusters were identified as:

- Cluster 1 – Traditional Forward Thinkers
- Cluster 2 – Satisfied, Free and Central
- Cluster 3 – Religious Traditionals
- Cluster 4 – Middling and Competitive

- Cluster 5 – Struggling Traditionals.

In the next section the clusters will be visualised from 1995-2009 along two dimensions that will be produced from canonical discriminant analysis of the variables in Table 24.

8.3 Visualisation

8.3.1 Visualisation Method

Canonical Discriminant Analysis does not require a strict assumption of normality. This allows analyses to accommodate dummy variables, as these are binary variables. Univariate normality is, however, recommended. For this reason, the numeric variables, of those that had been chosen in 8.1.1 and identified in Table 24, were ranked according to their value. The normal scores were then computed to help them appear more normally distributed (Blom, 1958). All empty dummy variables were deleted i.e. cases where a categorical variable has a category that has no entries. Canonical discriminant analysis was then carried out on these variables, with related cluster and country data for each of the years from 1995 to 2009.

A general rule of thumb in interpreting factors is to only consider correlations $> \pm$ 0.3 (Child, 2006). The pooled within class canonical structure removes between-class variability before computing the correlations. Variables with pooled within class correlations > 0.3 were

identified as being important in interpreting the resultant dimensions and these were identified for each of the years from 1995 to 2009.

The countries and their associated clusters were then plotted against the first two factors from the canonical discriminant analyses from 1995 to 2009.

8.3.2 Visualisation Results

The canonical discriminant analyses were similar between years - the variables that were significant in regards to interpreting the dimensions remaining fairly constant.

The first two dimensions were interpreted as follows:

1. Canonical variable 1 (x-axis when plotted)

Moving to the right along the x-axis, you would expect to see more Catholics and fewer Muslims within a country's population. If a country has been colonized it will likely be of Spanish origin. You would also expect to see an increasing proportion of people over 65.

For all variables mentioned in the canonical variable 2 description to follow, canonical variable 1 has a loading with an opposing sign. For example, if a variable is positively correlated with canonical variable 2, it will be negatively correlated with canonical variable 1. However these

correlations were not greater than the descriptive threshold of 0.30, as mentioned in 8.3.1.

This variable can be generalised as a progression from: Populations of younger Muslims, potentially through Secular, to Older Catholics.

2. Canonical variable 2 (y-axis when plotted)

Moving from the bottom to the top up the y-axis, you would expect to see an increasing birth rate, more young people in the population as a proportion of the whole, more men over 65 working and more women and men with no schooling. You would expect to see fewer countries classified as 'most free', shorter life expectancies, less schooling and less health expenditure per capita. You would also expect countries to have lower GDPs.

This variable can be roughly generalised as a progression from "Developed" countries - rich, free, educated and healthy, to countries who are "Developing" - dealing with poverty, low schooling and poor health.

The first factor from the canonical discriminant analysis accounts for at least half, and, during the years there are only three clusters, up to three quarters of the variation in the data. The second factor still has an important role, explaining around a quarter of the variation explained.

The visualisation of the results can be found at: http://goo.gl/qBwgdM

showing the individual countries and clusters evolving over time. The data driving the visualisation can be downloaded from the link, or a subset (each country, by cluster and year) can be found in Appendix 6.

When interpreting the axes, in this case named "Developed to developing" and "Young Muslims – Secular – Older Catholics", it can be helpful to make an analogy to the well-known Body Mass Index or BMI, which is an index based on weight and height. A person can arrive at a particular BMI with various combinations of weight and height. This is the case with the dimensions from the canonical discriminant analysis, there are various ways to arrive at a particular point on the dimension and there are many more factors involved than just the two in BMI. The labels are therefore a generalisation.

At the bottom left of the image Figure 33 is the play/pause button. Next to this is a small arrow (circled in red) that controls the speed of the visualisation. This should be set to the slowest setting for viewing enhancement.

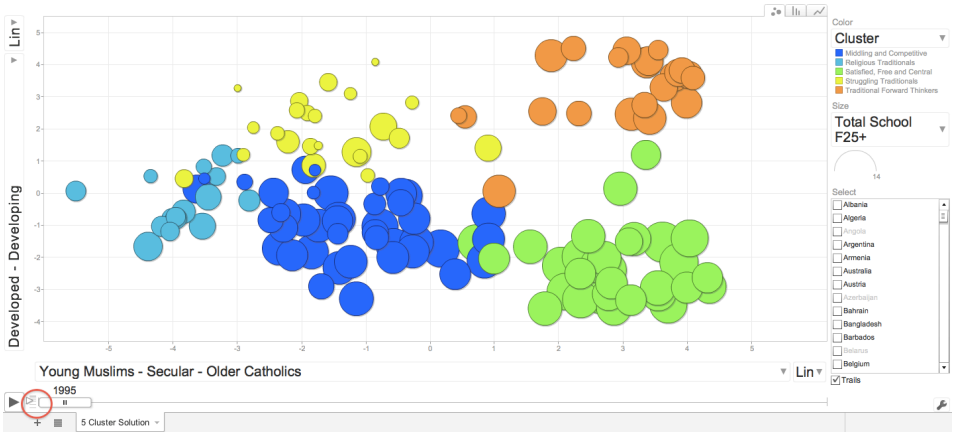


Figure 33: A screenshot from the final solution visualisation indicating the speed control button.

The bubble size has been set as "Total years schooling for women over 25" – a bigger bubble can be interpreted as more schooling. The bubble size can be set to any available variable that can be matched by country and by year. The Global Footprint would have been a very interesting bubble size variable, however, as mentioned previously, this is not available by year.

Before starting the visualisation, note that the smaller bubbles tend to be in the top left of the visualisation, and although the Traditional Forward Thinker cluster tends towards the Developing end of the Developed – Developing axis (DDA), the bubbles are reasonably large i.e. Women have a generally high level of schooling when compared to the Satisfied, Free and Central cluster. These countries also tend to be at a 'sustainable' level in regards to Global Footprint as shown in Figure 27.

The Middling and Competitive cluster can be seen to merge with the Satisfied, Free and Central cluster at times, as can the Religious Traditionals and Struggling Traditionals. The Traditional Forward Thinkers seem relatively stable over time.

It is interesting to highlight individual, or a couple of countries and watch their progression relative to others, over time.

8.3.3 Visualisation Summary

Canonical discriminant analyses using the most common variables as determined by the Genetic Algorithm approach in 8.1.1, produced consistent results over the years 1995 to 2009. The visualisation can be found here: http://goo.gl/qBwgdM

The visualisation plots the countries of the world (with available data) from 1995 to 2009 on two dimensions. These can be roughly described as: "Developed to Developing" and "Muslim with a relatively high proportion of young people (0-14s), through secular to Catholic with a relatively high proportion of older people (65+)".

9 Summary

Since the first half of last century, humans have used economic growth as a measure of progress. Not without merit, it has become the main indicator for the wellbeing of a nation. It is clear that if a nation does not have enough to sustain the needs of its population, then economic growth will bring increased life satisfaction, however, if a nation already has enough, then that is not necessarily true. Encouraging economic growth encourages consumption, and in 2013 the Global Footprint Network, a partner organisation of the New Economics Foundation, announced that the approximate day that our consumption exceeded our planet's ability to replenish was August 20th (Global Footprint Network, 2013).

There are a number of alternate progress measures already in existence, however, as with economic growth, the creator determines what is important, and therefore what components the measure contains – a top down approach.

In this work a bottom up approach was attempted, beginning in Section 4, with 290 variables, freely available in the public domain and associated with the areas considered important in (Stiglitz et al., 2009), in terms of future approaches to measuring our progress. In Section 5.3 a genetic algorithm was used in combination with forwards, backwards and stepwise regression, to select a subset of these variables, from each of the (Stiglitz et

al., 2009) areas in terms of their association with life satisfaction. Life satisfaction was used as a proxy for human flourishing.

This approach could have other applications, as many databases hold a large number of variables. For example, an analyst manually choosing which variables to include in a predictive model may not necessarily choose the best 'set'. If a genetic algorithm is used, the choice would be optimal in terms of whatever 'best' is, as decided by the analyst.

Next, a genetic algorithm was again used to determine if there are ways of existing in our world that are more conducive to flourishing, or living a 'better' life. Firstly, in Section 6.2.1 ("Approach 1"), the genetic algorithm variable selection method was used in conjunction with spectral clustering to obtain a 3-cluster solution. Spectral clustering gave a satisfactory 3-cluster solution, was relatively simple to implement and could be added to an analytical toolbox as an alternative to more traditional clustering techniques.

Secondly, in Section 6.2.2 ("Approach 2") a genetic algorithm, with a multiple component fitness function, was used to select not only the best subset of variables, but also the number of clusters and the best clustering method from a range of hierarchical agglomerative methods.

Until this point, the work had only been based on one year of data from 1995. In Sections 7.2 and 8.1 respectively, both Approach 1 and 2 were

extended into the future using an evolutionary method that aimed for the optimum balance between finding the best solution for the current year (snapshot quality), but not at the expense of too much deviation from the previous year (historical cost). Problems regarding continuity resulted in this approach being abandoned. However, this method would benefit from further investigation as it has the potential to provide superior results to the more traditional method of "scoring" future data based on an original model – an approach which has the potential to give poor snapshot quality.

Continuing in Section 8.1, the most common clustering method and model variables were identified from all years of Approach 2 results. These were then used for each year from 1996-2009 to ensure consistency between years. The number of clusters was permitted to be flexible from year to year e.g. 5 clusters in 1996, 3 clusters in 1997, and clusters were renamed when their centroid had very close association with the centroid of a cluster from the previous year. Despite losing the full snapshot quality that would have been obtained by allowing the genetic algorithm to choose a fresh clustering method and set of variables for each year, this approach is still better than 'scoring' future data based on an original model, as it allows a new solution while retaining continuity with previous years.

In Section 8.2 cluster profiling, or explanation, was carried out using:

1. Internal data: data used to build the clusters

2. External data: other data, not necessarily available for all years and all countries, such as the World Values Survey, Happy Planet Index. Further detail on the external data used can be found in Sections 8.2.1.1 to 8.2.1.3.

Full cluster descriptions can be found in Section 8.2.2.4. In general, a country in cluster 2 - "satisfied, free and central", has good health, good education, relatively good GDP. Further profiling, using external data, shows the highest life satisfaction in cluster 2. The countries with the least wealth have the lowest life satisfaction. Cluster 2 countries also have the highest global footprint, while cluster 1, "traditional forward thinker", countries have a relatively high life satisfaction and a relatively good sustainable footprint, as indicated by their placement in Figure 27. In the context of the research by (Myers, 2000) and (Kashdan, 2007) (described in 3.1), a possible research question raised by this result is "Why are people in cluster 1, the "traditional forward thinkers", relatively satisfied with life, while having less materially, and having a smaller footprint on the planet?"

There are similarities between country location on the WVS cultural map and cluster membership. A possible research question raised by this result is, "What impact does a particular set of values on determining a 'better' or 'worse' life?" If values do have an impact, then with a view to a

continued, satisfying existence on earth, further examination of the values of happy countries with low footprints would be of benefit.

The importance of variables relating to religion would benefit from further analysis. The most religious countries appear to be either relatively satisfied or alternatively, very unsatisfied. Are these variables related, or is the wealth of the country a confounding factor? Is there a difference between the impact of religious belief and religious practice on life satisfaction i.e. satisfaction gained from purpose versus satisfaction gained from practice?

In Section 8.3 a visualisation of the clusters over two dimensions derived from a canonical discriminant analysis on the internal data was produced, and can be found at: http://goo.gl/qBwgdM.

Figure 34 is a screen shot of the visualisation at 1995. The bubble size allows overlay of other measures of interest.

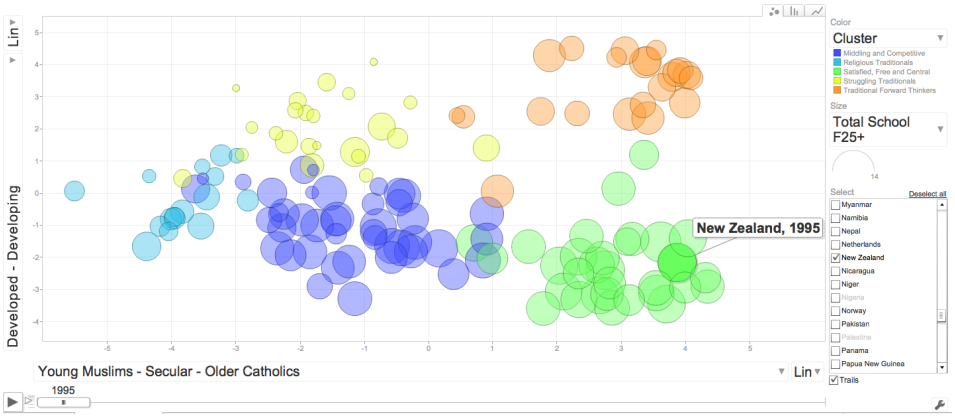


Figure 34: A screenshot of the visualisation. Bubble size indicates an overlayed variable of interest, in this case years at school for women over 25.

In Section 3.4 I stated:

I aim to use statistical techniques, in particular Genetic Algorithms, to define a measure of human progress. In contrast to the measures previously discussed, the components (or dimensions) of this measure will not be pre-defined, but rather will be defined statistically from a large number of variables that have been selected for their robustness and close statistical association with one of the areas (outlined earlier) in the (Stiglitz et al., 2009) report. This report has been a reference point for a number of the progress measures currently being worked on, including the United Kingdom's National Wellbeing work.
I aim to produce a three-dimensional (component) measure to enable ease of graphical visualisation over time (a fourth dimension) for as many countries in the world as there is data available. These dimensions, although statistically defined, should be linguistically interpretable, and should capture the essence of the areas defined in the (Stiglitz et al., 2009) report.

This work has met this statement. The visualisation is a quantitative display of 'Human Being', with countries in general living relatively better or worse lives with regards to Human Flourishing. The visualisation is a landscape, similar to that discussed by (Harris, 2010), containing the peaks and valleys of how and where we are over time, using data that have been collected regularly. There are components of flourishing in the

visualisation dimension explanations found in 8.3.2, but these are not easily shown on the plot. The profiles of the clusters provide a further view of progress, 'Human Flourishing', or a life well lived, but these again have to be read in conjunction with the visualisation.

Although this work has aimed to allow the data to deliver a result without deciding what variables should be contained in the measure (a bottom up, rather than top down approach), there was still some analytical decisions required, particularly with regard to the Genetic Algorithm fitness functions and interpretation of dimensions. Usually, in a project such as this, a team would be involved. At the end of three years work, the data are missing very recent years, but the methodology could certainly be extended to include new data.

Information for poorer countries was often missing which sometimes meant they were excluded. Although, as mentioned, information regarding our physical, immediate existence (e.g. schooling, health expenditure, GDP) was used in the construction of the clusters and visualisation, as it is easily available in consolidated form for many years. Some of the newer measures available in 2014 such as Global Footprint, or even Life Satisfaction (which are also potentially important factors in our continued progress), are not available each year, or do not involve all countries. As time goes on and data involving our emotional, or long term existence becomes more prevalent and accessible, these could be overlayed on a

visualisation such as that produced in this analysis, using the ‘bubble size’ i.e. In place of where the “Total Years School F25+” variable is now.

10 Personal Statement

The research question that preceded this work was: "Are there common dimensions to human flourishing measures and how do these dimensions relate to personal values?" In retrospect, there is a slight arrogance to the question – what is human flourishing, or a life well lived? For me there is no doubt that I would rather live in a "Satisfied, Free and Central" country than a "Struggling Traditional" country. I would expect to have a better opportunity to flourish. But what would a life well lived look like in a "Forward Thinking Traditional" country, where I may also have the opportunity to be reasonably satisfied and am likely to have less of a Global Footprint than in a "Satisfied, Free and Central" country.

I would rather live at the developed end of the Developed-Developing, y-axis, and in the secular section of the Young Muslim – Secular – Older Catholics, x-axis. Although there were enough data to somewhat answer the question "how do these dimensions relate to personal values?" in the profiles of the clusters, there is so much more that could be done if the data are available.

Perhaps, there are multiple ways in which to achieve a good life, but some are more conducive to the continued existence of our species.

On completion of this project, and in thinking of my own country, it is difficult for me to agree with our continued focus on economic growth.

New Zealanders are a reasonably satisfied people in general, but we are consuming resources at far more than a sustainable rate and the level of our income inequality is large. It is my belief that it is in these sorts of areas that we should be focusing our efforts if we want to progress, and achieve, as a nation, a life that could be considered flourishing.

When this project first began, I had no idea what an impact it would have personally. Professionally, I have gained a tremendous amount from this work. The simplicity of Genetic Algorithms and their potential application to analysing big data, the elegance of spectral clustering and finally alternative approaches to handling continuity on temporal data.

Personally, my worldview has shifted completely, and I have never felt more of a world citizen than I do now. It is no longer just about me, my family or even my country. I hope and intend to advocate that, rather than progress measures centred on our financial wealth, measures centred on the complete wellbeing or flourishing of our species, and the biosphere that is our home, will become more mainstream and important sooner rather than later.

Figure 35 is The Pale Blue Dot, a picture of Earth taken in 1990 at a distance of around 6 billion kilometres by the Voyager 1 space probe. The late astronomer Carl Sagan gave his interpretation of the photograph that consolidates my thoughts after completing this work:

"From this distant vantage point, the Earth might not seem of any particular interest. But for us, it's different. Consider again that dot. That's here. That's home. That's us. On it everyone you love, everyone you know, everyone you ever heard of, every human being who ever was, lived out their lives. The aggregate of our joy and suffering, thousands of confident religions, ideologies, and economic doctrines, every hunter and forager, every hero and coward, every creator and destroyer of civilization, every king and peasant, every young couple in love, every mother and father, hopeful child, inventor and explorer, every teacher of morals, every corrupt politician, every "superstar," every "supreme leader," every saint and sinner in the history of our species lived there – on a mote of dust suspended in a sunbeam.

The Earth is a very small stage in a vast cosmic arena. Think of the rivers of blood spilled by all those generals and emperors so that in glory and triumph they could become the momentary masters of a fraction of a dot. Think of the endless cruelties visited by the inhabitants of one corner of this pixel on the scarcely distinguishable inhabitants of some other corner. How frequent their misunderstandings, how eager they are to kill one another, how fervent their hatreds. Our posturings, our imagined self-importance, the delusion that we have some privileged position in the universe, are challenged by this point of pale light. Our planet is a lonely speck in the great enveloping cosmic dark. In our obscurity – in all this vastness – there is no hint that help will come from elsewhere to save us from ourselves.

The Earth is the only world known, so far, to harbour life. There is nowhere else, at least in the near future, to which our species could migrate. Visit, yes. Settle, not yet. Like it or not, for the moment, the Earth is where we make our stand. It has been said that astronomy is a humbling and character-building experience. There is perhaps no better demonstration of the folly of human conceits than this distant image of our tiny world. To me, it underscores our responsibility to deal more kindly with one another and to preserve and cherish the pale blue dot, the only home we've ever known."

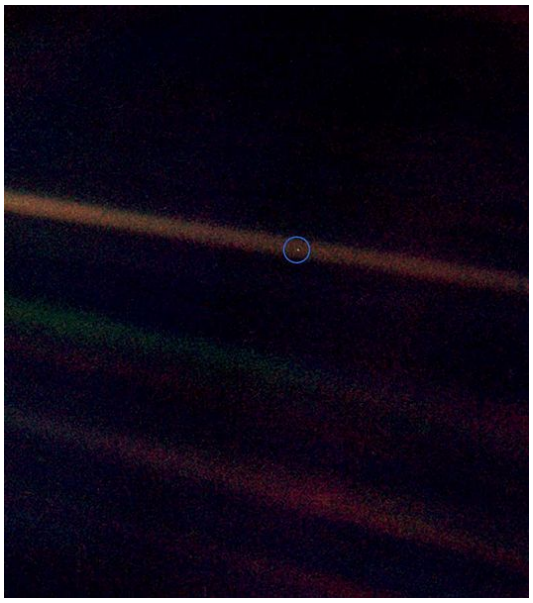


Figure 35: The pale blue dot – a picture of earth taken in 1990 at a distance of approximately 6 billion kilometres.

11 Bibliography

Abdallah, S., Michaelson, J., Shah, S., Stoll, L., & Marks, N. (2012). The Happy Planet Index: 2012 Report. A global index of sustainable well-being. The New Economics Foundation, UK.

AFH. (2010). Action for Happiness. Retrieved from http://www.actionforhappiness.org/about-us

Augustine, S. (1950). *The City of God*. New York: Fathers of the Church Inc.

Banerjee, B. (1994). How green is my value: exploring the relationship between environmentalism and materialism. *Advances in Consumer Research*, *21*, 147–152.

Barro, R., & Lee, J. (2010). *A New Data Set of Educational Attainment in the World 1950 - 2010* (NBER Working Paper No. 15902).

Baxter, M. J. (1995). Standardization and Transformation in Principal Component Analysis, with Applications to Archaeometry. *Journal of the Royal Statistical Society.*, *44*(4), 513–527.

Beal, D. J. (2005). SAS Code to Select the Best Multiple Linear Regression Model for Multivariate Data Using Information Criteria (Vol. Paper SA01_05). Portsmouth: South East SAS User Group.

Beyond GDP. (2007). Measuring Progress, True Wealth and the Well-being of Nations. Retrieved from http://www.beyond-gdp.eu/index.html

Blom, G. (1958). *Statistical estimates and transformed beta-variables*. New York: Wiley.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345:370.

Carroll, J. D. (1972). Individual Differences and Multidimensional Scaling. In *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences* (Vol. 1, pp. 105–153). New York: Seminar Press.

CASSE. (2003). Centre for the Advancement of Steady State Economy. Retrieved from http://www.steadystate.org

Centre for Bhutan Studies. (2012). *Centre for Bhutan Studies presents a film on GNH* (Vol. 1–3). Retrieved from http://www.grossnationalhappiness.com/multimedia/multimedia_p2/

Centre for Bhutan Studies. (n.d.). The GNH Index:What is it. Retrieved from http://www.grossnationalhappiness.com/articles/

Centre For Bhutan Studies. (n.d.). Gross National Happiness Index Explained in Detail. Retrieved from http://www.grossnationalhappiness.com/docs/GNH/PDFs/Sabina_Alkire_method.pdf

Chakrabarti, D., Kumar, R., & Tomkins, A. (2006). Evolutionary Clustering. Presented at the Knowledge Discovery and Data

Mining 2006. Retrieved from /~deepay/mywww/papers/kdd06-evolutionary.pdf

Child, D. (2006). *The Essentials of Factor Analysis* (3rd ed.). Continuum Publishing Group.

Chu, R., Duling, D., & Thompson, W. (2007). Best Practices for Managing Predictive Models in a Production Environment (Vol. SAS02). Presented at the Midwest Sas Users Group, Des Moines, Iowa. Retrieved from http://www.mwsug.org/proceedings/2007/saspres/MWSUG-2007-SAS02.pdf

Chung, F. (1997). *Spectral graph theory* (CBMS Regional Conference Series in Mathematics, Vol. 92). Washington: Conference Board of the MAthematical Sciences.

Cummins, R. A. (2000). Personal Income and Subjective Wellbeing: A review. *Journal of Happiness Studies*, *1*, 133–158.

Dasgupta, S., & Ng, V. (2010). Mining Clustering Dimensions. Presented at the The 27th International Conference on Machine Learning, Haifa, Israel. Retrieved from http://icml2010.haifa.il.ibm.com/papers/582.pdf

Diener, E. (1998). Subjective Wellbeing is Essential to Wellbeing. *Psychological Enquiry*, *9*(1), 33–37.

Edelstein, L. (1967). *The Idea of Progress in Classical Antiquity*. Baltimore: John Hopkins Press.

Fantom, N. (2014). How We Do Open Data. Retrieved from http://blogs.worldbank.org/opendata/how-we-do-open-data-1-choosing-development-indicators

Fodor, I. K. (2002). *A Survey of Dimension Reduction Techniques* (LLNL Technical Report). Lawrence Livermore National Laboratory.

Freedom House. (2012). *2012 Freedom in the World*. Retrieved from http://www.freedomhouse.org/report-types/freedom-world

Galbraith, J. (2010). Estimated Household Income Inequality Data Set (EHII). University of Texas Inequality Project. Retrieved from http://utip.gov.utexas.edu/data.html

Global Footprint Network. (2003). World Footprint: Do we fit on the planet? Retrieved from http://www.footprintnetwork.org/en/index.php/GFN/page/world_footprint/

Global Footprint Network. (2013). *Earth Overshoot Day 2013*. Retrieved from http://www.footprintnetwork.org/images/article_uploads/EarthOvershootDay_2013_PR_General.pdf

Goodall, C. (1983). M-Estimators of Location:An Outline of Theory. In *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, Ltd.

Grefenstette, J. (1986). Optimisation of Control Parameters for Genetic Algorithms. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-16*(1).

Guyon, I. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hadnagy, C. (2011). *Social Engineering: The art of human hacking*. Indianapolis, Indiana: Wiley Publishing.

Hagerty, M. R. (2003). Wealth and Happiness Revisited: Growing National Income Does Go with Greater Happiness. *Social Indicators Research*, (70), 243–255.

Hair, J., F., Anderson, R., E., Tatham, R., L., & Black, W., C. (1995). Stage Three: Assumptions in Canonical Correlation. In *Multivariate Data Analysis With Readings* (Fourth, p. 332). Prentice-Hall Inc.

Harris, S. (2010). *The Moral Landscape: How Science Can Determine Human Values*. Free Press.

Holland, J. (1993). *Adaptation in Natural And Artificial Systems* (2nd ed.). Massachusetts: Massachusetts Institute of Technology.

Hollander, M., Wolfe, D. A., & Chicken, E. (1973). *Non-Parametric Statistical Methods*. Wiley Publishing.

Human Development Report Office. (2010). About Human Development. Retrieved from http://hdr.undp.org/en/humandev/

Iglewicz, B. (1983). Robust scale estimators and confidence intervals for location. In *Understanding Robust and Exploratory Data Analysis* (pp. 404–431). John Wiley & Sons, Ltd.

Inglehart, R., & Welzel, C. (2010). Changing Mass Priorities: The link between modernization and democracy. *Perspectives on Politics*, *8*(2), 551–567.

International Labour Office. (1996a). LABORSTA Internet. Retrieved from http://laborsta.ilo.org/

International Labour Office. (1996b). Sources and methods: Labor Statistics. Retrieved from http://laborsta.ilo.org/applv8/data/SSMe.html

International Labour Organization. (1996, 2014). Retrieved from http://www.ilo.org/global/about-the-ilo/history/lang--en/index.htm

Istanbul Declaration. (2007). World Forum on Statistics, Knowledge and Policy. Retrieved from http://www.oecd.org/dataoecd/45/33/29130123.pdf

Jackson, T. (2009). *Prosperity Without Growth: Economics for a Finite Planet.* London: Earthscan.

Jain, A. K. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR) 31.3*, 264–323.

Kashdan, T. B. (2007). Materialism and Diminished Wellbeing. *Journal of Social and Clinical Psychology*, (26), 521–539.

Kharoufeh, J. P., & Goulias, Konstadinos G. (2002). Nonparametric identification of daily activity durations using kernel density estimators. *Transportation Research*, *Part B*(36), 59–82.

Lance, G. N., & Williams, W. T. (1967). A general theory of classification sorting strategies 1. Hierarchical systems. *The Computer Journal*, *9*(4), 373–380.

Matieny, P. (2000). The Psychodynamics of Meaning and Action for a Sustainable Future. *Futures*, (32), 339–360.

McDonough, W. (2005). *William McDonough on Cradle to*. USA. Retrieved from http://www.ted.com/talks/william_mcdonough_on_cradle_to_cradle_design.html

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Ltd.

Medrano, J. D. (2005). Managing Weights and Population weights. Banco de Datos ASEP / JDS. Retrieved from http://www.jdsurvey.net/jds/jdsurveyActualidad.jsp?Idioma=I&SeccionTexto=0405

Milligan, G. W. (1989). A Study of the Beta-Flexible Clustering Method. *Multivariate Behavioral Research*, *24*(2), 163–176.

Myers, D. G. (2000). The funds, friends and faith of Happy People. *American Psychologist*, *55*(1), 56–67.

Naldi, M. C., de Carvalho, A. C. P. L. F., Campello, R. J. G. B., & Hruschka, E. R. (2007). Genetic Clustering for Data Mining. In *Soft Computing for Knowledge Discovery and Data Mining* (pp. 113–132). Springer.

Narayanan, A., & Watts, D. (1996). Exact methods in the NPAR1WAY procedure. In *Proceedings of the Twenty-First Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc.

New Economics Foundation. (2006). Happy Planet Index. Retrieved from http://www.happyplanetindex.org

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On Spectral Clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, *2*, 849–856.

Noether, G. E. (1986). Why Kendall Tau? In *Best of Teaching Statistics*. University of Connecticut. Retrieved from http://www.rsscse-edu.org.uk/tsj/bts/noether/text.html

OECD. (2011). Better Life Index. Retrieved from http://oecdbetterlifeindex.org/#/

OFDA/CRED. (1988). *International Disaster Database*. Brussels, Belgium. Retrieved from www.emdat.be

Office of National Statistics. (2011). *Measuring National Wellbeing: Measuring What Matters.* Newport: Office of National Statistics, National Wellbeing Team.

Osborne, J., & Costello, A. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research & Evaluation*, *9*(11). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=11

Reiss, E. (2010). Using SAS Proc Cluster to Determine University Benchmarking Peers. Retrieved from http://analytics.ncsu.edu/sesug/2010/SDA10.Reiss.pdf

Rowe, J. (2008, June). Our Phony economy. *Harper's Magazine*, 17–24.

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Sachs, J. D. (2012). *World Happiness Report*. Columbia: Columbia University, The Earth Institute.

Sagiy, L., & Schwartz, S. (2000). Value priorities and subjective well-being: direct relations and congruity effects. *European Journal of Social Psychology*, *30*(2), 177–198.

SERI. (1999). Sustainable Europe Research Institute. Retrieved from http://seria.at/about/

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and MAchine Intelligence*, *22*(8), 888–905.

Sorensen, T. (1948). A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons. *Kongelige Danske Videnskabernes Selskab*, *5*, 1–34.

Statistics New Zealand. (2009). New Zealand General Social Survey. Retrieved from http://www.stats.govt.nz/browse_for_stats/people_and_communities/Well-being/nzgss-info-releases.aspx

Stiglitz, J. E., Sen, A., & Fitoussi, J.-P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress (Commission of the Government of France)*. Paris.

Symons, M. J. (1981). Clustering Criteria and Multivariate Normal Mixtures. *Biometrics*, *37*, 35–43.

Teorell, J. (2011). The Quality of Government Dataset. Retrieved from http://www.qog.pol.gu.se

The Carbon Trust. (2012, March). Carbon footprinting: The next step to reducing your emissions. The Carbon Trust. Retrieved from http://www.carbontrust.com/media/44869/j7912_ctv043_carbon_footprinting_aw_interactive.pdf

Turgo, A. R. J. (1973). *Turgot on Progress, Sociology and Economics*. Cambridge: University Press.

Ujjwal, M., & Sanghamitra, B. (2000). Genetic algorithm-based clustering technique. *Pattern Recognition*, *33*(9), 14550–1465.

UNEP. (2006). Environmental Data Explorer. Retrieved from http://geodata.grid.unep.ch/results.php

United Nations. (2000, September 8). 55/2. United Nations Millennium Declaration. Retrieved from http://www.un.org/millennium/declaration/ares552e.htm

Vinterbo, S. (1999). A Genetic Algorithm to Select Variables in Logistic Regression. *Journal of the American Medical Informatics Association 1999 Symposium Supplement*, 984–988.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416.

Wagstaff, K. et al. (2001). Constrained k-means clustering with background knowledge. (pp. 577–584). Presented at the Machine Learning-International Workshop Then Conference.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*, 263–244.

Weiss, Y. (1999). Segmentation using eigenvectors: A unifying view. *The Proceedings of the Seventh IEEE International Conference On.*, *2*.

well-being, n. (2015, March). *OED Online*. Oxford University Press. Retrieved from http://www.oed.com/view/Entry/227050?redirectedFrom=wellbeing

Wilkinson, R., & Pickett, K. (2010). *The Spirit Level: Why Equality is Better for Everyone*. London: Penguin Books.

Wilson, E. G. (2008). *Against Happiness: In Praise of Melancholy*. United States: Farrar, Straus and Giroux.

World Bank. (2011). World Bank Data. Retrieved from http://data.worldbank.org/data-catalog/research-datasets-analytical-tools

World Bank Group. (n.d.). World Development Indicators. Retrieved from http://data.worldbank.org/data-catalog/world-development-indicators

World Values Survey. (2008). *Values Change the World*. Retrieved from http://www.worldvaluessurvey.org/wvs/articles/folder_published/article_base_110/files/WVSbrochure6-2008_11.pdf

Xiang, T., & Gong, S. (2008). Spectral clustering with eigenvector selection. *Pattern Recognition*, *41*(2), 1012–1029.

Xu, K. S., Kliger, M., & Hero III, A. O. (2011). Adaptive evolutionary clustering. *Data Mining and Knowledge Discovery*, *1*(33).

Zhang, W. J., & Xu, L. (2001). Comparison of Different Methods for Variable Selection. *Analytica Chimica Acta*, *446*(1-2), 475–481.

12 Appendix 1

1. Aid effectiveness – 15 indicators selected	
Aid effectiveness is the impact that aid has in reducing poverty and inequality, increasing growth, building capacity, and accelerating achievement of the Millennium Development Goals set by the international community. Indicators here cover aid received as well as progress in reducing poverty and improving education, health, and other measures of human welfare	
INDICATOR_CODE	INDICATOR_NAME
DT.NFL.IFAD.CD	Net official flows from UN agencies, IFAD (current US$)
DT.NFL.UNAI.CD	Net official flows from UN agencies, UNAIDS (current US$)
DT.NFL.UNCF.CD	Net official flows from UN agencies, UNICEF (current US$)
DT.NFL.UNCR.CD	Net official flows from UN agencies, UNHCR (current US$)
DT.NFL.UNDP.CD	Net official flows from UN agencies, UNDP (current US$)
DT.NFL.UNFP.CD	Net official flows from UN agencies, UNFPA (current US$)
DT.NFL.UNRW.CD	Net official flows from UN agencies, UNRWA (current US$)
DT.NFL.UNTA.CD	Net official flows from UN agencies, UNTA (current US$)
DT.NFL.WFPG.CD	Net official flows from UN agencies, WFP (current US$)
DT.ODA.ALLD.CD	Net official development assistance and official aid received (current US$)
DT.ODA.ALLD.KD	Net official development assistance and official aid received (constant 2008 US$)
DT.ODA.OATL.CD	Net official aid received (current US$)
DT.ODA.OATL.KD	Net official aid received (constant 2008 US$)
DT.ODA.ODAT.CD	Net official development assistance received (current US$)
DT.ODA.ODAT.GI.ZS	Net ODA received (% of gross capital formation)
DT.ODA.ODAT.GN.ZS	Net ODA received (% of GNI)
DT.ODA.ODAT.KD	Net official development assistance received (constant 2008 US$)
DT.ODA.ODAT.MP.ZS	Net ODA received (% of imports of goods and services)

DT.ODA.ODAT.PC.ZS	Net ODA received per capita (current US$)
DT.ODA.ODAT.XP.ZS	Net ODA received (% of central government expense)
EN.ATM.CO2E.PC	CO2 emissions (metric tons per capita)
IT.CEL.SETS.P2	Mobile cellular subscriptions (per 100 people)
SE.ENR.PRSC.FM.ZS	Ratio of girls to boys in primary and secondary education (%)
SE.PRM.CMPT.ZS	Primary completion rate, total (% of relevant age group)
SG.GEN.PARL.ZS	Proportion of seats held by women in national parliaments (%)
SH.DYN.AIDS.ZS	Prevalence of HIV, total (% of population ages 15-49)
SH.DYN.MORT	Mortality rate, under-5 (per 1,000)
SH.STA.ACSN	Improved sanitation facilities (% of population with access)
SH.STA.ANVC.ZS	Pregnant women receiving prenatal care (%)
SH.STA.MALN.ZS	Malnutrition prevalence, weight for age (% of children under 5)
SH.STA.MMRT	Maternal mortality ratio (modeled estimate, per 100,000 live births)
SH.TBS.INCD	Incidence of tuberculosis (per 100,000 people)
SI.DST.FRST.20	Income share held by lowest 20%
SL.EMP.INSV.FE.ZS	Share of women employed in the nonagricultural sector (% of total nonagricultural employment)
SL.EMP.VULN.ZS	Vulnerable employment, total (% of total employment)
SM.POP.NETM	Net migration
SP.DYN.CONU.ZS	Contraceptive prevalence (% of women ages 15-49)
SP.DYN.LE00.FE.IN	Life expectancy at birth, female (years)
SP.DYN.LE00.MA.IN	Life expectancy at birth, male (years)
SP.MTR.1519.ZS	Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)
SP.POP.TOTL.FE.ZS	Population, female (% of total)
2. Economic policy and external debt – 24 indicators,	

Economic growth is central to economic development. When national income grows, real people benefit. While there is no known formula for stimulating economic growth, data can help policy-makers better understand their countries' economic situations and guide any work toward improvement. Data here covers measures of economic growth, such as gross domestic product (GDP) and gross national income (GNI). It also includes indicators representing factors known to be relevant to economic growth, such as capital stock, employment, investment, savings, consumption, government spending, imports, and exports.Debt statistics provide a detailed picture of debt stocks and flows of developing countries. Data presented as part of the Quarterly External Debt Statistics takes a closer look at the external debt of high-income countries and emerging markets to enable a more complete understanding of global financial flows. The Quarterly Public Sector Debt database provides further data on public sector valuation methods, debt instruments, and clearly defined tiers of debt for central, state and local government, as well as extra-budgetary agencies and funds. Data are gathered from national statistical organizations and central banks as well as by various major multilateral institutions and World Bank staff.

INDICATOR_CODE	INDICATOR_NAME
BG.GSR.NFSV.GD.ZS	Trade in services (% of GDP)
BM.GSR.CMCP.ZS	Communications, computer, etc. (% of service imports, BoP)
BM.GSR.FCTY.CD	Income payments (BoP, current US$)
BM.GSR.GNFS.CD	Imports of goods and services (BoP, current US$)
BM.GSR.MRCH.CD	Goods imports (BoP, current US$)
BM.GSR.NFSV.CD	Service imports (BoP, current US$)
BM.GSR.ROYL.CD	Royalty and license fees, payments (BoP, current US$)
BM.GSR.TOTL.CD	Imports of goods, services and income (BoP, current US$)
BM.GSR.TRAN.ZS	Transport services (% of service imports, BoP)
BM.GSR.TRVL.ZS	Travel services (% of service imports, BoP)
BM.TRF.PRVT.CD	Private current transfers, payments (BoP, current US$)
BN.CAB.XOKA.CD	Current account balance (BoP, current US$)
BN.CAB.XOKA.GD.ZS	Current account balance (% of GDP)
BN.GSR.FCTY.CD	Net income (BoP, current US$)
BN.GSR.GNFS.CD	Net trade in goods and services (BoP, current US$)
BN.GSR.MRCH.CD	Net trade in goods (BoP, current US$)
BN.KAC.EOMS.CD	Net errors and omissions, adjusted (BoP, current US$)
BN.KLT.PRVT.CD	Private capital flows, total (BoP, current US$)
BN.KLT.PRVT.GD.ZS	Private capital flows, total (% of GDP)

BN.RES.INCL.CD	Changes in net reserves (BoP, current US$)
BN.TRF.CURR.CD	Net current transfers (BoP, current US$)
BN.TRF.KOGT.CD	Net capital account (BoP, current US$)
BX.GSR.CMCP.ZS	Communications, computer, etc. (% of service exports, BoP)
BX.GSR.FCTY.CD	Income receipts (BoP, current US$)
BX.GSR.GNFS.CD	Exports of goods and services (BoP, current US$)
BX.GSR.MRCH.CD	Goods exports (BoP, current US$)
BX.GSR.NFSV.CD	Service exports (BoP, current US$)
BX.GSR.ROYL.CD	Royalty and license fees, receipts (BoP, current US$)
BX.GSR.TOTL.CD	Exports of goods, services and income (BoP, current US$)
BX.GSR.TRAN.ZS	Transport services (% of service exports, BoP)
BX.GSR.TRVL.ZS	Travel services (% of service exports, BoP)
BX.KLT.DINV.CD.WD	Foreign direct investment, net inflows (BoP, current US$)
BX.PEF.TOTL.CD.WD	Portfolio equity, net inflows (BoP, current US$)
BX.TRF.CURR.CD	Current transfers, receipts (BoP, current US$)
BX.TRF.PWKR.CD.DT	Workers' remittances and compensation of employees, received (current US$)
DC.DAC.AUSL.CD	Net bilateral aid flows from DAC donors, Australia (current US$)
DC.DAC.AUTL.CD	Net bilateral aid flows from DAC donors, Austria (current US$)
DC.DAC.BELL.CD	Net bilateral aid flows from DAC donors, Belgium (current US$)
DC.DAC.CANL.CD	Net bilateral aid flows from DAC donors, Canada (current US$)
DC.DAC.CECL.CD	Net bilateral aid flows from DAC donors, European Union institutions (current US$)
DC.DAC.CHEL.CD	Net bilateral aid flows from DAC donors, Switzerland (current US$)
DC.DAC.DEUL.CD	Net bilateral aid flows from DAC donors, Germany (current US$)
DC.DAC.DNKL.CD	Net bilateral aid flows from DAC donors, Denmark (current US$)
DC.DAC.ESPL.CD	Net bilateral aid flows from DAC donors, Spain (current US$)
DC.DAC.FINL.CD	Net bilateral aid flows from DAC donors, Finland (current US$)
DC.DAC.FRAL.CD	Net bilateral aid flows from DAC donors, France (current US$)
DC.DAC.GBRL.CD	Net bilateral aid flows from DAC donors, United Kingdom (current US$)
DC.DAC.GRCL.CD	Net bilateral aid flows from DAC donors, Greece (current US$)
DC.DAC.IRLL.CD	Net bilateral aid flows from DAC donors, Ireland (current US$)

DC.DAC.ITAL.CD	Net bilateral aid flows from DAC donors, Italy (current US$)
DC.DAC.JPNL.CD	Net bilateral aid flows from DAC donors, Japan (current US$)
DC.DAC.LUXL.CD	Net bilateral aid flows from DAC donors, Luxembourg (current US$)
DC.DAC.NLDL.CD	Net bilateral aid flows from DAC donors, Netherlands (current US$)
DC.DAC.NORL.CD	Net bilateral aid flows from DAC donors, Norway (current US$)
DC.DAC.NZLL.CD	Net bilateral aid flows from DAC donors, New Zealand (current US$)
DC.DAC.PRTL.CD	Net bilateral aid flows from DAC donors, Portugal (current US$)
DC.DAC.SWEL.CD	Net bilateral aid flows from DAC donors, Sweden (current US$)
DC.DAC.TOTL.CD	Net bilateral aid flows from DAC donors, Total (current US$)
DC.DAC.USAL.CD	Net bilateral aid flows from DAC donors, United States (current US$)
DT.DOD.DECT.CD	External debt stocks, total (DOD, current US$)
DT.DOD.DECT.GN.ZS	External debt stocks (% of GNI)
DT.DOD.DIMF.CD	Use of IMF credit (DOD, current US$)
DT.DOD.DLXF.CD	External debt stocks, long-term (DOD, current US$)
DT.DOD.DPNG.CD	External debt stocks, private nonguaranteed (PNG) (DOD, current US$)
DT.DOD.DPPG.CD	External debt stocks, public and publicly guaranteed (PPG) (DOD, current US$)
DT.DOD.DSTC.CD	External debt stocks, short-term (DOD, current US$)
DT.DOD.DSTC.IR.ZS	Short-term debt (% of total reserves)
DT.DOD.DSTC.XP.ZS	Short-term debt (% of exports of goods, services and income)
DT.DOD.DSTC.ZS	Short-term debt (% of total external debt)
DT.DOD.MWBG.CD	IBRD loans and IDA credits (DOD, current US$)
DT.DOD.PVLX.CD	Present value of external debt (current US$)
DT.DOD.PVLX.EX.ZS	Present value of external debt (% of exports of goods, services and income)
DT.DOD.PVLX.GN.ZS	Present value of external debt (% of GNI)
DT.NFL.BLAT.CD	Net financial flows, bilateral (NFL, current US$)
DT.NFL.IAEA.CD	Net official flows from UN agencies, IAEA (current US$)
DT.NFL.IMFC.CD	Net financial flows, IMF concessional (NFL, current US$)
DT.NFL.IMFN.CD	Net financial flows, IMF nonconcessional (NFL, current US$)
DT.NFL.MIBR.CD	Net financial flows, IBRD (NFL, current US$)
DT.NFL.MIDA.CD	Net financial flows, IDA (NFL, current US$)

DT.NFL.MLAT.CD	Net financial flows, multilateral (NFL, current US$)
DT.NFL.MOTH.CD	Net financial flows, others (NFL, current US$)
DT.NFL.PCBO.CD	Commercial banks and other lending (PPG + PNG) (NFL, current US$)
DT.NFL.RDBC.CD	Net financial flows, RDB concessional (NFL, current US$)
DT.NFL.RDBN.CD	Net financial flows, RDB nonconcessional (NFL, current US$)
DT.NFL.UNEC.CD	Net official flows from UN agencies, UNECE (current US$)
DT.NFL.WHOL.CD	Net official flows from UN agencies, WHO (current US$)
DT.ODA.ALLD.CD	Net official development assistance and official aid received (current US$)
DT.ODA.ODAT.CD	Net official development assistance received (current US$)
DT.ODA.ODAT.GN.ZS	Net ODA received (% of GNI)
DT.ODA.ODAT.PC.ZS	Net ODA received per capita (current US$)
DT.TDS.DECT.CD	Debt service on external debt, total (TDS, current US$)
DT.TDS.DECT.EX.ZS	Total debt service (% of exports of goods, services and income)
DT.TDS.DECT.GN.ZS	Total debt service (% of GNI)
DT.TDS.DPPF.XP.ZS	Debt service (PPG and IMF only, % of exports, excluding workers' remittances)
DT.TDS.DPPG.CD	Debt service on external debt, public and publicly guaranteed (PPG) (TDS, current US$)
DT.TDS.DPPG.GN.ZS	Public and publicly guaranteed debt service (% of GNI)
DT.TDS.DPPG.XP.ZS	Public and publicly guaranteed debt service (% of exports, excluding workers' remittances)
DT.TDS.MLAT.CD	Multilateral debt service (TDS, current US$)
DT.TDS.MLAT.PG.ZS	Multilateral debt service (% of public and publicly guaranteed debt service)
FI.RES.TOTL.CD	Total reserves (includes gold, current US$)
FI.RES.TOTL.DT.ZS	Total reserves (% of total external debt)
FI.RES.TOTL.MO	Total reserves in months of imports
FI.RES.XGLD.CD	Total reserves minus gold (current US$)
FP.CPI.TOTL.ZG	Inflation, consumer prices (annual %)
GC.BAL.CASH.CN	Cash surplus/deficit (current LCU)
GC.BAL.CASH.GD.ZS	Cash surplus/deficit (% of GDP)
GC.DOD.TOTL.CN	Central government debt, total (current LCU)
GC.DOD.TOTL.GD.ZS	Central government debt, total (% of GDP)

GC.REV.XGRT.GD.ZS	Revenue, excluding grants (% of GDP)
GC.XPN.COMP.CN	Compensation of employees (current LCU)
GC.XPN.COMP.ZS	Compensation of employees (% of expense)
NE.CON.GOVT.CD	General government final consumption expenditure (current US$)
NE.CON.GOVT.CN	General government final consumption expenditure (current LCU)
NE.CON.GOVT.KD	General government final consumption expenditure (constant 2000 US$)
NE.CON.GOVT.KD.ZG	General government final consumption expenditure (annual % growth)
NE.CON.GOVT.KN	General government final consumption expenditure (constant LCU)
NE.CON.GOVT.ZS	General government final consumption expenditure (% of GDP)
NE.CON.PETC.CD	Household final consumption expenditure, etc. (current US$)
NE.CON.PETC.CN	Household final consumption expenditure, etc. (current LCU)
NE.CON.PETC.KD	Household final consumption expenditure, etc. (constant 2000 US$)
NE.CON.PETC.KD.ZG	Household final consumption expenditure, etc. (annual % growth)
NE.CON.PETC.KN	Household final consumption expenditure, etc. (constant LCU)
NE.CON.PETC.ZS	Household final consumption expenditure, etc. (% of GDP)
NE.CON.PRVT.CD	Household final consumption expenditure (current US$)
NE.CON.PRVT.CN	Household final consumption expenditure (current LCU)
NE.CON.PRVT.KD	Household final consumption expenditure (constant 2000 US$)
NE.CON.PRVT.KD.ZG	Household final consumption expenditure (annual % growth)
NE.CON.PRVT.KN	Household final consumption expenditure (constant LCU)
NE.CON.PRVT.PC.KD	Household final consumption expenditure per capita (constant 2000 US$)
NE.CON.PRVT.PC.KD.ZG	Household final consumption expenditure per capita growth (annual %)
NE.CON.PRVT.PP.CD	Household final consumption expenditure, PPP (current international $)
NE.CON.PRVT.PP.KD	Household final consumption expenditure, PPP (constant 2005 international $)
NE.CON.TETC.CD	Final consumption expenditure, etc. (current US$)
NE.CON.TETC.CN	Final consumption expenditure, etc. (current LCU)
NE.CON.TETC.KD	Final consumption expenditure, etc. (constant 2000 US$)
NE.CON.TETC.KD.ZG	Final consumption expenditure, etc. (annual % growth)
NE.CON.TETC.KN	Final consumption expenditure, etc. (constant LCU)
NE.CON.TETC.ZS	Final consumption expenditure, etc. (% of GDP)

NE.CON.TOTL.CD	Final consumption expenditure (current US$)
NE.CON.TOTL.CN	Final consumption expenditure (current LCU)
NE.CON.TOTL.KD	Final consumption expenditure (constant 2000 US$)
NE.CON.TOTL.KN	Final consumption expenditure (constant LCU)
NE.DAB.DEFL.ZS	Gross national expenditure deflator (base year varies by country)
NE.DAB.TOTL.CD	Gross national expenditure (current US$)
NE.DAB.TOTL.CN	Gross national expenditure (current LCU)
NE.DAB.TOTL.KD	Gross national expenditure (constant 2000 US$)
NE.DAB.TOTL.KN	Gross national expenditure (constant LCU)
NE.DAB.TOTL.ZS	Gross national expenditure (% of GDP)
NE.EXP.GNFS.CD	Exports of goods and services (current US$)
NE.EXP.GNFS.CN	Exports of goods and services (current LCU)
NE.EXP.GNFS.KD	Exports of goods and services (constant 2000 US$)
NE.EXP.GNFS.KD.ZG	Exports of goods and services (annual % growth)
NE.EXP.GNFS.KN	Exports of goods and services (constant LCU)
NE.EXP.GNFS.ZS	Exports of goods and services (% of GDP)
NE.GDI.FPRV.CN	Gross fixed capital formation, private sector (current LCU)
NE.GDI.FPRV.ZS	Gross fixed capital formation, private sector (% of GDP)
NE.GDI.FTOT.CD	Gross fixed capital formation (current US$)
NE.GDI.FTOT.CN	Gross fixed capital formation (current LCU)
NE.GDI.FTOT.KD	Gross fixed capital formation (constant 2000 US$)
NE.GDI.FTOT.KD.ZG	Gross fixed capital formation (annual % growth)
NE.GDI.FTOT.KN	Gross fixed capital formation (constant LCU)
NE.GDI.FTOT.ZS	Gross fixed capital formation (% of GDP)
NE.GDI.STKB.CD	Changes in inventories (current US$)
NE.GDI.STKB.CN	Changes in inventories (current LCU)
NE.GDI.STKB.KN	Changes in inventories (constant LCU)
NE.GDI.TOTL.CD	Gross capital formation (current US$)
NE.GDI.TOTL.CN	Gross capital formation (current LCU)
NE.GDI.TOTL.KD	Gross capital formation (constant 2000 US$)

NE.GDI.TOTL.KD.ZG	Gross capital formation (annual % growth)
NE.GDI.TOTL.KN	Gross capital formation (constant LCU)
NE.GDI.TOTL.ZS	Gross capital formation (% of GDP)
NE.IMP.GNFS.CD	Imports of goods and services (current US$)
NE.IMP.GNFS.CN	Imports of goods and services (current LCU)
NE.IMP.GNFS.KD	Imports of goods and services (constant 2000 US$)
NE.IMP.GNFS.KD.ZG	Imports of goods and services (annual % growth)
NE.IMP.GNFS.KN	Imports of goods and services (constant LCU)
NE.IMP.GNFS.ZS	Imports of goods and services (% of GDP)
NE.RSB.GNFS.CD	External balance on goods and services (current US$)
NE.RSB.GNFS.CN	External balance on goods and services (current LCU)
NE.RSB.GNFS.KN	External balance on goods and services (constant LCU)
NE.RSB.GNFS.ZS	External balance on goods and services (% of GDP)
NE.TRD.GNFS.ZS	Trade (% of GDP)
NV.AGR.TOTL.CD	Agriculture, value added (current US$)
NV.AGR.TOTL.CN	Agriculture, value added (current LCU)
NV.AGR.TOTL.KD	Agriculture, value added (constant 2000 US$)
NV.AGR.TOTL.KD.ZG	Agriculture, value added (annual % growth)
NV.AGR.TOTL.KN	Agriculture, value added (constant LCU)
NV.AGR.TOTL.ZS	Agriculture, value added (% of GDP)
NV.IND.MANF.CD	Manufacturing, value added (current US$)
NV.IND.MANF.CN	Manufacturing, value added (current LCU)
NV.IND.MANF.KD	Manufacturing, value added (constant 2000 US$)
NV.IND.MANF.KD.ZG	Manufacturing, value added (annual % growth)
NV.IND.MANF.KN	Manufacturing, value added (constant LCU)
NV.IND.MANF.ZS	Manufacturing, value added (% of GDP)
NV.IND.TOTL.CD	Industry, value added (current US$)
NV.IND.TOTL.CN	Industry, value added (current LCU)
NV.IND.TOTL.KD	Industry, value added (constant 2000 US$)
NV.IND.TOTL.KD.ZG	Industry, value added (annual % growth)

NV.IND.TOTL.KN	Industry, value added (constant LCU)
NV.IND.TOTL.ZS	Industry, value added (% of GDP)
NV.MNF.CHEM.ZS.UN	Chemicals (% of value added in manufacturing)
NV.MNF.FBTO.ZS.UN	Food, beverages and tobacco (% of value added in manufacturing)
NV.MNF.MTRN.ZS.UN	Machinery and transport equipment (% of value added in manufacturing)
NV.MNF.OTHR.ZS.UN	Other manufacturing (% of value added in manufacturing)
NV.MNF.TXTL.ZS.UN	Textiles and clothing (% of value added in manufacturing)
NV.SRV.TETC.CD	Services, etc., value added (current US$)
NV.SRV.TETC.CN	Services, etc., value added (current LCU)
NV.SRV.TETC.KD	Services, etc., value added (constant 2000 US$)
NV.SRV.TETC.KD.ZG	Services, etc., value added (annual % growth)
NV.SRV.TETC.KN	Services, etc., value added (constant LCU)
NV.SRV.TETC.ZS	Services, etc., value added (% of GDP)
NY.ADJ.AEDU.CD	Adjusted savings: education expenditure (current US$)
NY.ADJ.AEDU.GN.ZS	Adjusted savings: education expenditure (% of GNI)
NY.ADJ.DCO2.CD	Adjusted savings: carbon dioxide damage (current US$)
NY.ADJ.DCO2.GN.ZS	Adjusted savings: carbon dioxide damage (% of GNI)
NY.ADJ.DFOR.CD	Adjusted savings: net forest depletion (current US$)
NY.ADJ.DFOR.GN.ZS	Adjusted savings: net forest depletion (% of GNI)
NY.ADJ.DKAP.CD	Adjusted savings: consumption of fixed capital (current US$)
NY.ADJ.DKAP.GN.ZS	Adjusted savings: consumption of fixed capital (% of GNI)
NY.ADJ.DMIN.CD	Adjusted savings: mineral depletion (current US$)
NY.ADJ.DMIN.GN.ZS	Adjusted savings: mineral depletion (% of GNI)
NY.ADJ.DNGY.CD	Adjusted savings: energy depletion (current US$)
NY.ADJ.DNGY.GN.ZS	Adjusted savings: energy depletion (% of GNI)
NY.ADJ.DPEM.CD	Adjusted savings: particulate emission damage (current US$)
NY.ADJ.DPEM.GN.ZS	Adjusted savings: particulate emission damage (% of GNI)
NY.ADJ.DRES.GN.ZS	Adjusted savings: natural resources depletion (% of GNI)
NY.ADJ.ICTR.GN.ZS	Adjusted savings: gross savings (% of GNI)
NY.ADJ.NNAT.CD	Adjusted savings: net national savings (current US$)

NY.ADJ.NNAT.GN.ZS	Adjusted savings: net national savings (% of GNI)
NY.ADJ.NNTY.CD	Adjusted net national income (current US$)
NY.ADJ.NNTY.KD	Adjusted net national income (constant 2000 US$)
NY.ADJ.NNTY.KD.ZG	Adjusted net national income (annual % growth)
NY.ADJ.SVNG.CD	Adjusted net savings, including particulate emission damage (current US$)
NY.ADJ.SVNG.GN.ZS	Adjusted net savings, including particulate emission damage (% of GNI)
NY.ADJ.SVNX.CD	Adjusted net savings, excluding particulate emission damage (current US$)
NY.ADJ.SVNX.GN.ZS	Adjusted net savings, excluding particulate emission damage (% of GNI)
NY.EXP.CAPM.KN	Exports as a capacity to import (constant LCU)
NY.GDP.DEFL.KD.ZG	Inflation, GDP deflator (annual %)
NY.GDP.DEFL.ZS	GDP deflator (base year varies by country)
NY.GDP.DISC.CN	Discrepancy in expenditure estimate of GDP (current LCU)
NY.GDP.DISC.KN	Discrepancy in expenditure estimate of GDP (constant LCU)
NY.GDP.FCST.CD	Gross value added at factor cost (current US$)
NY.GDP.FCST.CN	Gross value added at factor cost (current LCU)
NY.GDP.FCST.KD	Gross value added at factor cost (constant 2000 US$)
NY.GDP.FCST.KN	Gross value added at factor cost (constant LCU)
NY.GDP.MKTP.CD	GDP (current US$)
NY.GDP.MKTP.CN	GDP (current LCU)
NY.GDP.MKTP.KD	GDP (constant 2000 US$)
NY.GDP.MKTP.KD.ZG	GDP growth (annual %)
NY.GDP.MKTP.KN	GDP (constant LCU)
NY.GDP.MKTP.PP.CD	GDP, PPP (current international $)
NY.GDP.MKTP.PP.KD	GDP, PPP (constant 2005 international $)
NY.GDP.PCAP.CD	GDP per capita (current US$)
NY.GDP.PCAP.KD	GDP per capita (constant 2000 US$)
NY.GDP.PCAP.KD.ZG	GDP per capita growth (annual %)
NY.GDP.PCAP.KN	GDP per capita (constant LCU)
NY.GDP.PCAP.PP.CD	GDP per capita, PPP (current international $)
NY.GDP.PCAP.PP.KD	GDP per capita, PPP (constant 2005 international $)

NY.GDS.TOTL.CD	Gross domestic savings (current US$)
NY.GDS.TOTL.CN	Gross domestic savings (current LCU)
NY.GDS.TOTL.KN	Gross domestic savings (constant LCU)
NY.GDS.TOTL.ZS	Gross domestic savings (% of GDP)
NY.GDY.TOTL.KD	Gross domestic income (constant 2000 US$)
NY.GDY.TOTL.KN	Gross domestic income (constant LCU)
NY.GNP.ATLS.CD	GNI, Atlas method (current US$)
NY.GNP.MKTP.CD	GNI (current US$)
NY.GNP.MKTP.CN	GNI (current LCU)
NY.GNP.MKTP.KD	GNI (constant US$)
NY.GNP.MKTP.KD.ZG	GNI growth (annual %)
NY.GNP.MKTP.KN	GNI (constant LCU)
NY.GNP.MKTP.PP.CD	GNI, PPP (current international $)
NY.GNP.PCAP.CD	GNI per capita, Atlas method (current US$)
NY.GNP.PCAP.KD	GNI per capita (constant 2000 US$)
NY.GNP.PCAP.KD.ZG	GNI per capita growth, constant 2000$ (annual %)
NY.GNP.PCAP.KN	GNI per capita (constant LCU)
NY.GNP.PCAP.PP.CD	GNI per capita, PPP (current international $)
NY.GNS.ICTR.CD	Gross savings (current US$)
NY.GNS.ICTR.CN	Gross savings (current LCU)
NY.GNS.ICTR.GN.ZS	Gross savings (% of GNI)
NY.GNS.ICTR.ZS	Gross savings (% of GDP)
NY.GNY.TOTL.KN	Gross national income (constant LCU)
NY.GSR.NFCY.CD	Net income from abroad (current US$)
NY.GSR.NFCY.CN	Net income from abroad (current LCU)
NY.GSR.NFCY.KN	Net income from abroad (constant LCU)
NY.TAX.NIND.CD	Net taxes on products (current US$)
NY.TAX.NIND.CN	Net taxes on products (current LCU)
NY.TAX.NIND.KN	Net taxes on products (constant LCU)
NY.TRF.NCTR.CD	Net current transfers from abroad (current US$)

NY.TRF.NCTR.CN	Net current transfers from abroad (current LCU)
NY.TRF.NCTR.KN	Net current transfers from abroad (constant LCU)
NY.TTF.GNFS.KN	Terms of trade adjustment (constant LCU)
PA.NUS.PPP	PPP conversion factor, GDP (LCU per international $)
PA.NUS.PPPC.RF	PPP conversion factor (GDP) to market exchange rate ratio
PA.NUS.PRVT.PP	PPP conversion factor, private consumption (LCU per international $)
3. Education – 6 indicators,	
Education is one of the most powerful instruments for reducing poverty and inequality and lays a foundation for sustained economic growth. The World Bank compiles data on education inputs, participation, efficiency, and outcomes. Data on education are compiled by the United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute for Statistics from official responses to surveys and from reports provided by education authorities in each country.	
INDICATOR_CODE	INDICATOR_NAME
SE.ADT.1524.LT.FE.ZS	Literacy rate, youth female (% of females ages 15-24)
SE.ADT.1524.LT.MA.ZS	Literacy rate, youth male (% of males ages 15-24)
SE.ADT.1524.LT.ZS	Literacy rate, youth total (% of people ages 15-24)
SE.ADT.LITR.FE.ZS	Literacy rate, adult female (% of females ages 15 and above)
SE.ADT.LITR.MA.ZS	Literacy rate, adult male (% of males ages 15 and above)
SE.ADT.LITR.ZS	Literacy rate, adult total (% of people ages 15 and above)
SE.ENR.PRIM.FM.ZS	Ratio of female to male primary enrollment (%)
SE.ENR.PRSC.FM.ZS	Ratio of girls to boys in primary and secondary education (%)
SE.ENR.SECO.FM.ZS	Ratio of female to male secondary enrollment (%)
SE.ENR.TERT.FM.ZS	Ratio of female to male tertiary enrollment (%)
SE.PRE.ENRR	School enrollment, preprimary (% gross)
SE.PRE.ENRR.FE	School enrollment, preprimary, female (% gross)
SE.PRE.ENRR.MA	School enrollment, preprimary, male (% gross)
SE.PRM.AGES	Primary school starting age (years)
SE.PRM.CMPT.FE.ZS	Primary completion rate, female (% of relevant age group)
SE.PRM.CMPT.MA.ZS	Primary completion rate, male (% of relevant age group)
SE.PRM.CMPT.ZS	Primary completion rate, total (% of relevant age group)
SE.PRM.DURS	Primary education, duration (years)

SE.PRM.ENRL	Primary education, pupils
SE.PRM.ENRL.FE.ZS	Primary education, pupils (% female)
SE.PRM.ENRL.TC.ZS	Pupil-teacher ratio, primary
SE.PRM.ENRR	School enrollment, primary (% gross)
SE.PRM.ENRR.FE	School enrollment, primary, female (% gross)
SE.PRM.ENRR.MA	School enrollment, primary, male (% gross)
SE.PRM.GINT.FE.ZS	Gross intake rate in grade 1, female (% of relevant age group)
SE.PRM.GINT.MA.ZS	Gross intake rate in grade 1, male (% of relevant age group)
SE.PRM.GINT.ZS	Gross intake rate in grade 1, total (% of relevant age group)
SE.PRM.NENR	School enrollment, primary (% net)
SE.PRM.NENR.FE	School enrollment, primary, female (% net)
SE.PRM.NENR.MA	School enrollment, primary, male (% net)
SE.PRM.NINT.FE.ZS	Net intake rate in grade 1, female (% of official school-age population)
SE.PRM.NINT.MA.ZS	Net intake rate in grade 1, male (% of official school-age population)
SE.PRM.NINT.ZS	Net intake rate in grade 1 (% of official school-age population)
SE.PRM.PRIV.ZS	School enrollment, primary, private (% of total primary)
SE.PRM.PRS5.FE.ZS	Persistence to grade 5, female (% of cohort)
SE.PRM.PRS5.MA.ZS	Persistence to grade 5, male (% of cohort)
SE.PRM.PRS5.ZS	Persistence to grade 5, total (% of cohort)
SE.PRM.PRSL.FE.ZS	Persistence to last grade of primary, female (% of cohort)
SE.PRM.PRSL.MA.ZS	Persistence to last grade of primary, male (% of cohort)
SE.PRM.PRSL.ZS	Persistence to last grade of primary, total (% of cohort)
SE.PRM.REPT.FE.ZS	Repeaters, primary, female (% of female enrollment)
SE.PRM.REPT.MA.ZS	Repeaters, primary, male (% of male enrollment)
SE.PRM.REPT.ZS	Repeaters, primary, total (% of total enrollment)
SE.PRM.TCAQ.FE.ZS	Trained teachers in primary education, female (% of female teachers)
SE.PRM.TCAQ.MA.ZS	Trained teachers in primary education, male (% of male teachers)
SE.PRM.TCAQ.ZS	Trained teachers in primary education (% of total teachers)
SE.PRM.TCHR	Primary education, teachers
SE.PRM.TCHR.FE.ZS	Primary education, teachers (% female)

SE.PRM.TENR	Total enrollment, primary (% net)
SE.PRM.TENR.FE	Total enrollment, primary, female (% net)
SE.PRM.TENR.MA	Total enrollment, primary, male (% net)
SE.PRM.UNER	Children out of school, primary
SE.PRM.UNER.FE	Children out of school, primary, female
SE.PRM.UNER.MA	Children out of school, primary, male
SE.SEC.AGES	Secondary school starting age (years)
SE.SEC.DURS	Secondary education, duration (years)
SE.SEC.ENRL	Secondary education, pupils
SE.SEC.ENRL.FE.ZS	Secondary education, pupils (% female)
SE.SEC.ENRL.GC	Secondary education, general pupils
SE.SEC.ENRL.GC.FE.ZS	Secondary education, general pupils (% female)
SE.SEC.ENRL.TC.ZS	Pupil-teacher ratio, secondary
SE.SEC.ENRL.VO	Secondary education, vocational pupils
SE.SEC.ENRL.VO.FE.ZS	Secondary education, vocational pupils (% female)
SE.SEC.ENRR	School enrollment, secondary (% gross)
SE.SEC.ENRR.FE	School enrollment, secondary, female (% gross)
SE.SEC.ENRR.MA	School enrollment, secondary, male (% gross)
SE.SEC.NENR	School enrollment, secondary (% net)
SE.SEC.NENR.FE	School enrollment, secondary, female (% net)
SE.SEC.NENR.MA	School enrollment, secondary, male (% net)
SE.SEC.PRIV.ZS	School enrollment, secondary, private (% of total secondary)
SE.SEC.PROG.FE.ZS	Progression to secondary school, female (%)
SE.SEC.PROG.MA.ZS	Progression to secondary school, male (%)
SE.SEC.PROG.ZS	Progression to secondary school (%)
SE.SEC.REPT.FE.ZS	Repeaters, secondary, female (% of female enrollment)
SE.SEC.REPT.MA.ZS	Repeaters, secondary, male (% of male enrollment)
SE.SEC.REPT.ZS	Repeaters, secondary, total (% of total enrollment)
SE.SEC.TCHR	Secondary education, teachers
SE.SEC.TCHR.FE	Secondary education, teachers, female

SE.SEC.TCHR.FE.ZS	Secondary education, teachers (% female)
SE.TER.ENRR	School enrollment, tertiary (% gross)
SE.TER.ENRR.FE	School enrollment, tertiary, female (% gross)
SE.TER.ENRR.MA	School enrollment, tertiary, male (% gross)
SE.XPD.PRIM.PC.ZS	Expenditure per student, primary (% of GDP per capita)
SE.XPD.SECO.PC.ZS	Expenditure per student, secondary (% of GDP per capita)
SE.XPD.TERT.PC.ZS	Expenditure per student, tertiary (% of GDP per capita)
SE.XPD.TOTL.GB.ZS	Public spending on education, total (% of government expenditure)
SE.XPD.TOTL.GD.ZS	Public spending on education, total (% of GDP)
4. Environment – 19 indicators,	
Natural and man-made environmental resources Ð fresh water, clean air, forests, grasslands, marine resources, and agro-ecosystems Ð provide sustenance and a foundation for social and economic development. The need to safeguard these resources crosses all borders. Today, the World Bank is one of the key promoters and financiers of environmental upgrading in the developing world. Data here cover forests, biodiversity, emissions, and pollution. Other indicators relevant to the environment are found under data pages for Agriculture & Rural Development, Energy & Mining, Infrastructure, and Urban Development.	
INDICATOR_CODE	INDICATOR_NAME
AG.LND.FRST.K2	Forest area (sq. km)
AG.LND.FRST.ZS	Forest area (% of land area)
AG.PRD.CREL.MT	Cereal production (metric tons)
EE.BOD.CGLS.ZS	Water pollution, clay and glass industry (% of total BOD emissions)
EE.BOD.CHEM.ZS	Water pollution, chemical industry (% of total BOD emissions)
EE.BOD.FOOD.ZS	Water pollution, food industry (% of total BOD emissions)
EE.BOD.MTAL.ZS	Water pollution, metal industry (% of total BOD emissions)
EE.BOD.OTHR.ZS	Water pollution, other industry (% of total BOD emissions)
EE.BOD.PAPR.ZS	Water pollution, paper and pulp industry (% of total BOD emissions)
EE.BOD.TOTL.KG	Organic water pollutant (BOD) emissions (kg per day)
EE.BOD.TXTL.ZS	Water pollution, textile industry (% of total BOD emissions)
EE.BOD.WOOD.ZS	Water pollution, wood industry (% of total BOD emissions)
EE.BOD.WRKR.KG	Organic water pollutant (BOD) emissions (kg per day per worker)
EG.ELC.FOSL.ZS	Electricity production from oil, gas and coal sources (% of total)

EN.ATM.CO2E.EG.ZS	CO2 intensity (kg per kg of oil equivalent energy use)
EN.ATM.CO2E.KD.GD	CO2 emissions (kg per 2000 US$ of GDP)
EN.ATM.CO2E.KT	CO2 emissions (kt)
EN.ATM.CO2E.PC	CO2 emissions (metric tons per capita)
EN.ATM.CO2E.PP.GD	CO2 emissions (kg per PPP $ of GDP)
EN.ATM.CO2E.PP.GD.KD	CO2 emissions (kg per 2005 PPP $ of GDP)
EN.ATM.CO2E.SF.ZS	CO2 emissions from solid fuel consumption (% of total)
EN.ATM.GHGO.KT.CE	Other greenhouse gas emissions, HFC, PFC and SF6 (thousand metric tons of CO2 equivalent)
EN.ATM.HFCG.KT.CE	HFC gas emissions (thousand metric tons of CO2 equivalent)
EN.ATM.METH.AG.KT.CE	Agricultural methane emissions (thousand metric tons of CO2 equivalent)
EN.ATM.METH.AG.ZS	Agricultural methane emissions (% of total)
EN.ATM.METH.EG.KT.CE	Methane emissions in energy sector (thousand metric tons of CO2 equivalent)
EN.ATM.METH.EG.ZS	Energy related methane emissions (% of total)
EN.ATM.METH.KT.CE	Methane emissions (kt of CO2 equivalent)
EN.ATM.NOXE.AG.KT.CE	Agricultural nitrous oxide emissions (thousand metric tons of CO2 equivalent)
EN.ATM.NOXE.AG.ZS	Agricultural nitrous oxide emissions (% of total)
EN.ATM.NOXE.EG.KT.CE	Nitrous oxide emissions in energy sector (thousand metric tons of CO2 equivalent)
EN.ATM.NOXE.EI.ZS	Nitrous oxide emissions in industrial and energy processes (% of total nitrous oxide emissions)
EN.ATM.NOXE.IN.KT.CE	Industrial nitrous oxide emissions (thousand metric tons of CO2 equivalent)
EN.ATM.NOXE.KT.CE	Nitrous oxide emissions (thousand metric tons of CO2 equivalent)
EN.ATM.PFCG.KT.CE	PFC gas emissions (thousand metric tons of CO2 equivalent)
EN.ATM.SF6G.KT.CE	SF6 gas emissions (thousand metric tons of CO2 equivalent)
EN.BIR.THRD.NO	Bird species, threatened
EN.FSH.THRD.NO	Fish species, threatened
EN.HPT.THRD.NO	Plant species (higher), threatened
EN.MAM.THRD.NO	Mammal species, threatened
ER.BDV.TOTL.XQ	GEF benefits index for biodiversity (0 = no biodiversity potential to 100 = maximum)

ER.GDP.FWTL.M3.KD	Water productivity, total (constant 2000 US$ GDP per cubic meter of total freshwater withdrawal)
ER.LND.PTLD.ZS	Nationally protected areas (% of total area)
ER.MRN.PTMR.ZS	Marine protected areas (% of total surface area)
NY.ADJ.AEDU.CD	Adjusted savings: education expenditure (current US$)
NY.ADJ.AEDU.GN.ZS	Adjusted savings: education expenditure (% of GNI)
NY.ADJ.DCO2.CD	Adjusted savings: carbon dioxide damage (current US$)
NY.ADJ.DCO2.GN.ZS	Adjusted savings: carbon dioxide damage (% of GNI)
NY.ADJ.DFOR.CD	Adjusted savings: net forest depletion (current US$)
NY.ADJ.DFOR.GN.ZS	Adjusted savings: net forest depletion (% of GNI)
NY.ADJ.DKAP.CD	Adjusted savings: consumption of fixed capital (current US$)
NY.ADJ.DKAP.GN.ZS	Adjusted savings: consumption of fixed capital (% of GNI)
NY.ADJ.DMIN.CD	Adjusted savings: mineral depletion (current US$)
NY.ADJ.DMIN.GN.ZS	Adjusted savings: mineral depletion (% of GNI)
NY.ADJ.DNGY.CD	Adjusted savings: energy depletion (current US$)
NY.ADJ.DNGY.GN.ZS	Adjusted savings: energy depletion (% of GNI)
NY.ADJ.DPEM.CD	Adjusted savings: particulate emission damage (current US$)
NY.ADJ.DPEM.GN.ZS	Adjusted savings: particulate emission damage (% of GNI)
NY.ADJ.ICTR.GN.ZS	Adjusted savings: gross savings (% of GNI)
NY.ADJ.NNAT.CD	Adjusted savings: net national savings (current US$)
NY.ADJ.NNAT.GN.ZS	Adjusted savings: net national savings (% of GNI)
NY.ADJ.SVNG.CD	Adjusted net savings, including particulate emission damage (current US$)
NY.ADJ.SVNG.GN.ZS	Adjusted net savings, including particulate emission damage (% of GNI)
NY.ADJ.SVNX.CD	Adjusted net savings, excluding particulate emission damage (current US$)
NY.ADJ.SVNX.GN.ZS	Adjusted net savings, excluding particulate emission damage (% of GNI)
NY.GDP.COAL.RT.ZS	Coal rents (% of GDP)
NY.GDP.FRST.RT.ZS	Forest rents (% of GDP)
NY.GDP.MINR.RT.ZS	Mineral rents (% of GDP)
NY.GDP.NGAS.RT.ZS	Natural gas rents (% of GDP)
NY.GDP.PETR.RT.ZS	Oil rents (% of GDP)
NY.GDP.TOTL.RT.ZS	Total natural resources rents (% of GDP)

5. Financial sector – 6 indicators,	
An economy's financial markets are critical to its overall development. Banking systems and stock markets enhance growth, the main factor in poverty reduction. Strong financial systems provide reliable and accessible information that lowers transaction costs, which in turn bolsters resource allocation and economic growth. Indicators here include the size and liquidity of stock markets; the accessibility, stability, and efficiency of financial systems; and international migration and workers\ remittances, which affect growth and social welfare in both sending and receiving countries.	
INDICATOR_CODE	INDICATOR_NAME
BM.KLT.DINV.GD.ZS	Foreign direct investment, net outflows (% of GDP)
BM.TRF.PWKR.CD.DT	Workers' remittances and compensation of employees, paid (current US$)
BN.KLT.DINV.CD	Foreign direct investment, net (BoP, current US$)
BN.KLT.PTXL.CD	Portfolio investment, excluding LCFAR (BoP, current US$)
BX.KLT.DINV.CD.WD	Foreign direct investment, net inflows (BoP, current US$)
BX.KLT.DINV.WD.GD.ZS	Foreign direct investment, net inflows (% of GDP)
BX.PEF.TOTL.CD.WD	Portfolio equity, net inflows (BoP, current US$)
BX.TRF.PWKR.CD	Workers' remittances, receipts (BoP, current US$)
BX.TRF.PWKR.CD.DT	Workers' remittances and compensation of employees, received (current US$)
BX.TRF.PWKR.DT.GD.ZS	Workers' remittances and compensation of employees, received (% of GDP)
CM.FIN.INTL.GD.ZS	Financing via international capital markets (gross inflows, % of GDP)
CM.MKT.INDX.ZG	S&P Global Equity Indices (annual % change)
CM.MKT.LCAP.CD	Market capitalization of listed companies (current US$)
CM.MKT.LCAP.GD.ZS	Market capitalization of listed companies (% of GDP)
CM.MKT.LDOM.NO	Listed domestic companies, total
CM.MKT.TRAD.CD	Stocks traded, total value (current US$)
CM.MKT.TRAD.GD.ZS	Stocks traded, total value (% of GDP)
CM.MKT.TRNR	Stocks traded, turnover ratio (%)
DT.NFL.BOND.CD	Portfolio investment, bonds (PPG + PNG) (NFL, current US$)
FB.AST.LOAN.CB.P3	Loan accounts, commercial banks (per 1,000 adults)
FB.AST.LOAN.CO.P3	Loan accounts, cooperatives (per 1,000 adults)
FB.AST.LOAN.MF.P3	Loan accounts, microfinance institutions (per 1,000 adults)
FB.AST.LOAN.SF.P3	Loan accounts, specialized state financial institutions (per 1,000 adults)

FB.AST.NPER.ZS	Bank nonperforming loans to total gross loans (%)
FB.ATM.TOTL.P5	Automated teller machines (ATMs) (per 100,000 adults)
FB.BNK.BRCH.CB.P5	Branches, commercial banks (per 100,000 adults)
FB.BNK.BRCH.CO.P5	Branches, cooperatives (per 100,000 adults)
FB.BNK.BRCH.MF.P5	Branches, microfinance institutions (per 100,000 adults)
FB.BNK.BRCH.SF.P5	Branches, specialized state financial institutions (per 100,000 adults)
FB.BNK.CAPA.ZS	Bank capital to assets ratio (%)
FB.LBL.DDPT.CB.P3	Deposit accounts, commercial banks (per 1,000 adults)
FB.LBL.DDPT.CO.P3	Deposit accounts, cooperatives (per 1,000 adults)
FB.LBL.DDPT.MF.P3	Deposit accounts, microfinance institutions (per 1,000 adults)
FB.LBL.DDPT.SF.P3	Deposit accounts, specialized state financial institutions (per 1,000 adults)
FB.POS.TOTL.P5	Point-of-sale terminals (per 100,000 adults)
FD.RES.LIQU.AS.ZS	Bank liquid reserves to bank assets ratio (%)
FI.RES.TOTL.CD	Total reserves (includes gold, current US$)
FM.AST.CGOV.ZG.M3	Claims on central government (annual growth as % of broad money)
FM.AST.DOMO.ZG.M3	Claims on other sectors of the domestic economy (annual growth as % of broad money)
FM.AST.DOMS.CN	Net domestic credit (current LCU)
FM.AST.NFRG.CN	Net foreign assets (current LCU)
FM.AST.PRVT.ZG.M3	Claims on private sector (annual growth as % of broad money)
FM.LBL.BMNY.CN	Broad money (current LCU)
FM.LBL.BMNY.GD.ZS	Broad money (% of GDP)
FM.LBL.BMNY.IR.ZS	Broad money to total reserves ratio
FM.LBL.BMNY.ZG	Broad money growth (annual %)
FM.LBL.MONY.CN	Money (current LCU)
FM.LBL.MQMY.CN	Money and quasi money (M2) (current LCU)
FM.LBL.MQMY.GD.ZS	Money and quasi money (M2) as % of GDP
FM.LBL.MQMY.IR.ZS	Money and quasi money (M2) to total reserves ratio
FM.LBL.MQMY.ZG	Money and quasi money growth (annual %)
FM.LBL.QMNY.CN	Quasi money (current LCU)
FP.CPI.TOTL	Consumer price index (2005 = 100)

FP.WPI.TOTL	Wholesale price index (2005 = 100)
FR.INR.DPST	Deposit interest rate (%)
FR.INR.LEND	Lending interest rate (%)
FR.INR.LNDP	Interest rate spread (lending rate minus deposit rate, %)
FR.INR.RINR	Real interest rate (%)
FR.INR.RISK	Risk premium on lending (prime rate minus treasury bill rate, %)
FS.AST.CGOV.GD.ZS	Claims on central government, etc. (% GDP)
FS.AST.DOMO.GD.ZS	Claims on other sectors of the domestic economy (% of GDP)
FS.AST.DOMS.GD.ZS	Domestic credit provided by banking sector (% of GDP)
FS.LBL.LIQU.GD.ZS	Liquid liabilities (M3) as % of GDP
FS.LBL.QLIQ.GD.ZS	Quasi-liquid liabilities (% of GDP)
IC.CRD.INFO.XQ	Credit depth of information index (0=low to 6=high)
IC.CRD.PRVT.ZS	Private credit bureau coverage (% of adults)
IC.CRD.PUBL.ZS	Public credit registry coverage (% of adults)
IC.LGL.CRED.XQ	Strength of legal rights index (0=weak to 10=strong)
PA.NUS.ATLS	DEC alternative conversion factor (LCU per US$)
PA.NUS.FCRF	Official exchange rate (LCU per US$, period average)
PX.REX.REER	Real effective exchange rate index (2005 = 100)
SM.EMI.TERT.ZS	Emigration rate of tertiary educated (% of total tertiary educated population)
SM.POP.NETM	Net migration
SM.POP.TOTL	International migrant stock, total
SM.POP.TOTL.ZS	International migrant stock (% of population)
6. Health – 25 indicators,	
Improving health is central to the Millennium Development Goals, and the public sector is the main provider of health care in developing countries. To reduce inequities, many countries have emphasized primary health care, including immunization, sanitation, access to safe drinking water, and safe motherhood initiatives. Data here cover health systems, disease prevention, reproductive health, nutrition, and population dynamics. Data are from the United Nations Population Division, World Health Organization, United Nations Children's Fund, the Joint United Nations Programme on HIV/AIDS, and various other sources.	
INDICATOR_CODE	INDICATOR_NAME
SH.CON.1524.FE.ZS	Condom use, population ages 15-24, female (% of females ages 15-24)

SH.CON.1524.MA.ZS	Condom use, population ages 15-24, male (% of males ages 15-24)
SH.DYN.AIDS.FE.ZS	Female adults with HIV (% of population ages 15+ with HIV)
SH.DYN.AIDS.ZS	Prevalence of HIV, total (% of population ages 15-49)
SH.DYN.CHLD.FE	Mortality rate, female child (per 1,000 female children age one)
SH.DYN.CHLD.MA	Mortality rate, male child (per 1,000 male children age one)
SH.DYN.MORT	Mortality rate, under-5 (per 1,000)
SH.HIV.0014	Children (0-14) living with HIV
SH.HIV.1524.FE.ZS	Prevalence of HIV, female (% ages 15-24)
SH.HIV.1524.MA.ZS	Prevalence of HIV, male (% ages 15-24)
SH.IMM.IDPT	Immunization, DPT (% of children ages 12-23 months)
SH.IMM.MEAS	Immunization, measles (% of children ages 12-23 months)
SH.MED.BEDS.ZS	Hospital beds (per 1,000 people)
SH.MED.CMHW.P3	Community health workers (per 1,000 people)
SH.MED.NUMW.P3	Nurses and midwives (per 1,000 people)
SH.MED.PHYS.ZS	Physicians (per 1,000 people)
SH.MLR.NETS.ZS	Use of insecticide-treated bed nets (% of under-5 population)
SH.MLR.TRET.ZS	Children with fever receiving antimalarial drugs (% of children under age 5 with fever)
SH.MMR.RISK	Lifetime risk of maternal death (1 in: rate varies by country)
SH.MMR.RISK.ZS	Lifetime risk of maternal death (%)
SH.PRV.SMOK.FE	Smoking prevalence, females (% of adults)
SH.PRV.SMOK.MA	Smoking prevalence, males (% of adults)
SH.STA.ACSN	Improved sanitation facilities (% of population with access)
SH.STA.ACSN.RU	Improved sanitation facilities, rural (% of rural population with access)
SH.STA.ACSN.UR	Improved sanitation facilities, urban (% of urban population with access)
SH.STA.ANVC.ZS	Pregnant women receiving prenatal care (%)
SH.STA.ARIC.ZS	ARI treatment (% of children under 5 taken to a health provider)
SH.STA.BFED.ZS	Exclusive breastfeeding (% of children under 6 months)
SH.STA.BRTC.ZS	Births attended by skilled health staff (% of total)
SH.STA.BRTW.ZS	Low-birthweight babies (% of births)
SH.STA.MALN.ZS	Malnutrition prevalence, weight for age (% of children under 5)

SH.STA.MMRT	Maternal mortality ratio (modeled estimate, per 100,000 live births)
SH.STA.MMRT.NE	Maternal mortality ratio (national estimate, per 100,000 live births)
SH.STA.ORCF.ZS	Diarrhea treatment (% of children under 5 receiving oral rehydration and continued feeding)
SH.STA.OWGH.ZS	Prevalence of overweight (% of children under 5)
SH.STA.STNT.ZS	Malnutrition prevalence, height for age (% of children under 5)
SH.STA.WAST.ZS	Prevalence of wasting (% of children under 5)
SH.TBS.CURE.ZS	Tuberculosis treatment success rate (% of registered cases)
SH.TBS.DTEC.ZS	Tuberculosis case detection rate (%, all forms)
SH.TBS.INCD	Incidence of tuberculosis (per 100,000 people)
SH.VAC.TTNS.ZS	Newborns protected against tetanus (%)
SH.VST.OUTP	Outpatient visits per capita
SH.XPD.EXTR.ZS	External resources for health (% of total expenditure on health)
SH.XPD.OOPC.TO.ZS	Out-of-pocket health expenditure (% of total expenditure on health)
SH.XPD.OOPC.ZS	Out-of-pocket health expenditure (% of private expenditure on health)
SH.XPD.PCAP	Health expenditure per capita (current US$)
SH.XPD.PCAP.PP.KD	Health expenditure per capita, PPP (constant 2005 international $)
SH.XPD.PRIV.ZS	Health expenditure, private (% of GDP)
SH.XPD.PUBL	Health expenditure, public (% of total health expenditure)
SH.XPD.PUBL.GX.ZS	Health expenditure, public (% of government expenditure)
SH.XPD.PUBL.ZS	Health expenditure, public (% of GDP)
SH.XPD.TOTL.ZS	Health expenditure, total (% of GDP)
SN.ITK.DEFC.ZS	Prevalence of undernourishment (% of population)
SN.ITK.DPTH	Depth of hunger (kilocalories per person per day)
SN.ITK.SALT.ZS	Consumption of iodized salt (% of households)
SN.ITK.VITA.ZS	Vitamin A supplementation coverage rate (% of children ages 6-59 months)
SP.ADO.TFRT	Adolescent fertility rate (births per 1,000 women ages 15-19)
SP.DTH.INFR.ZS	Completeness of infant death reporting (% of reported infant deaths to estimated infant deaths)
SP.DTH.REPT.ZS	Completeness of total death reporting (% of reported total deaths to estimated total deaths)

SP.DYN.AMRT.FE	Mortality rate, adult, female (per 1,000 female adults)
SP.DYN.AMRT.MA	Mortality rate, adult, male (per 1,000 male adults)
SP.DYN.CBRT.IN	Birth rate, crude (per 1,000 people)
SP.DYN.CDRT.IN	Death rate, crude (per 1,000 people)
SP.DYN.CONU.ZS	Contraceptive prevalence (% of women ages 15-49)
SP.DYN.IMRT.IN	Mortality rate, infant (per 1,000 live births)
SP.DYN.LE00.FE.IN	Life expectancy at birth, female (years)
SP.DYN.LE00.IN	Life expectancy at birth, total (years)
SP.DYN.LE00.MA.IN	Life expectancy at birth, male (years)
SP.DYN.TFRT.IN	Fertility rate, total (births per woman)
SP.DYN.TO65.FE.ZS	Survival to age 65, female (% of cohort)
SP.DYN.TO65.MA.ZS	Survival to age 65, male (% of cohort)
SP.DYN.WFRT	Wanted fertility rate (births per woman)
SP.HOU.FEMA.ZS	Female headed households (% of households with a female head)
SP.MTR.1519.ZS	Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)
SP.POP.0014.TO.ZS	Population ages 0-14 (% of total)
SP.POP.1564.TO.ZS	Population ages 15-64 (% of total)
SP.POP.65UP.TO.ZS	Population ages 65 and above (% of total)
SP.POP.DPND	Age dependency ratio (% of working-age population)
SP.POP.DPND.OL	Age dependency ratio, old (% of working-age population)
SP.POP.DPND.YG	Age dependency ratio, young (% of working-age population)
SP.POP.GROW	Population growth (annual %)
SP.POP.TOTL	Population, total
SP.POP.TOTL.FE.ZS	Population, female (% of total)
SP.REG.BRTH.RU.ZS	Completeness of birth registration, rural (%)
SP.REG.BRTH.UR.ZS	Completeness of birth registration, urban (%)
SP.REG.BRTH.ZS	Completeness of birth registration (%)
SP.UWT.TFRT	Unmet need for contraception (% of married women ages 15-49)
7. Labour and social protection – 8 indicators,	

The supply of labor available in an economy includes people who are employed, those who are unemployed but seeking work, and first-time job-seekers. Not everyone who works is included: unpaid workers, family workers, and students are often omitted, while some countries do not count members of the armed forces. Data on labor and employment are compiled by the International Labour Organization (ILO) from labor force surveys, censuses, establishment censuses and surveys, and administrative records such as employment exchange registers and unemployment insurance schemes.

INDICATOR_CODE	INDICATOR_NAME
SL.AGR.0714.FE.ZS	Child employment in agriculture, female (% of female economically active children ages 7-14)
SL.AGR.0714.MA.ZS	Child employment in agriculture, male (% of male economically active children ages 7-14)
SL.AGR.0714.ZS	Child employment in agriculture (% of economically active children ages 7-14)
SL.AGR.EMPL.FE.ZS	Employees, agriculture, female (% of female employment)
SL.AGR.EMPL.MA.ZS	Employees, agriculture, male (% of male employment)
SL.AGR.EMPL.ZS	Employment in agriculture (% of total employment)
SL.EMP.1524.SP.FE.ZS	Employment to population ratio, ages 15-24, female (%)
SL.EMP.1524.SP.MA.ZS	Employment to population ratio, ages 15-24, male (%)
SL.EMP.1524.SP.ZS	Employment to population ratio, ages 15-24, total (%)
SL.EMP.MPYR.FE.ZS	Employers, female (% of employment)
SL.EMP.MPYR.MA.ZS	Employers, male (% of employment)
SL.EMP.MPYR.ZS	Employers, total (% of employment)
SL.EMP.SELF.FE.ZS	Self-employed, female (% of females employed)
SL.EMP.SELF.MA.ZS	Self-employed, male (% of males employed)
SL.EMP.SELF.ZS	Self-employed, total (% of total employed)
SL.EMP.TOTL.SP.FE.ZS	Employment to population ratio, 15+, female (%)
SL.EMP.TOTL.SP.MA.ZS	Employment to population ratio, 15+, male (%)
SL.EMP.TOTL.SP.ZS	Employment to population ratio, 15+, total (%)
SL.EMP.VULN.FE.ZS	Vulnerable employment, female (% of female employment)
SL.EMP.VULN.MA.ZS	Vulnerable employment, male (% of male employment)
SL.EMP.VULN.ZS	Vulnerable employment, total (% of total employment)
SL.EMP.WORK.FE.ZS	Wage and salaried workers, female (% of females employed)

SL.EMP.WORK.MA.ZS	Wage and salary workers, male (% of males employed)
SL.EMP.WORK.ZS	Wage and salaried workers, total (% of total employed)
SL.FAM.WORK.FE.ZS	Contributing family workers, female (% of females employed)
SL.FAM.WORK.MA.ZS	Contributing family workers, male (% of males employed)
SL.FAM.WORK.ZS	Contributing family workers, total (% of total employed)
SL.GDP.PCAP.EM.KD	GDP per person employed (constant 1990 PPP $)
SL.IND.EMPL.FE.ZS	Employees, industry, female (% of female employment)
SL.IND.EMPL.MA.ZS	Employees, industry, male (% of male employment)
SL.IND.EMPL.ZS	Employment in industry (% of total employment)
SL.MNF.0714.FE.ZS	Child employment in manufacturing, female (% of female economically active children ages 7-14)
SL.MNF.0714.MA.ZS	Child employment in manufacturing, male (% of male economically active children ages 7-14)
SL.MNF.0714.ZS	Child employment in manufacturing (% of economically active children ages 7-14)
SL.MNF.WAGE.FM	Ratio of female to male wages in manufacturing
SL.SRV.0714.FE.ZS	Child employment in services, female (% of female economically active children ages 7-14)
SL.SRV.0714.MA.ZS	Child employment in services, male (% of male economically active children ages 7-14)
SL.SRV.0714.ZS	Child employment in services (% of economically active children ages 7-14)
SL.SRV.EMPL.FE.ZS	Employees, services, female (% of female employment)
SL.SRV.EMPL.MA.ZS	Employees, services, male (% of male employment)
SL.SRV.EMPL.ZS	Employment in services (% of total employment)
SL.TLF.CACT.FE.ZS	Labor participation rate, female (% of female population ages 15+)
SL.TLF.CACT.MA.ZS	Labor participation rate, male (% of male population ages 15+)
SL.TLF.CACT.ZS	Labor participation rate, total (% of total population ages 15+)
SL.TLF.PART.FE.ZS	Part time employment, female (% of total female employment)
SL.TLF.PART.MA.ZS	Part time employment, male (% of total male employment)
SL.TLF.PART.TL.FE.ZS	Part time employment, female (% of total part time employment)
SL.TLF.PART.ZS	Part time employment, total (% of total employment)

SL.TLF.PRIM.FE.ZS	Labor force with primary education, female (% of female labor force)
SL.TLF.PRIM.MA.ZS	Labor force with primary education, male (% of male labor force)
SL.TLF.PRIM.ZS	Labor force with primary education (% of total)
SL.TLF.SECO.FE.ZS	Labor force with secondary education, female (% of female labor force)
SL.TLF.SECO.MA.ZS	Labor force with secondary education, male (% of male labor force)
SL.TLF.SECO.ZS	Labor force with secondary education (% of total)
SL.TLF.TERT.FE.ZS	Labor force with tertiary education, female (% of female labor force)
SL.TLF.TERT.MA.ZS	Labor force with tertiary education, male (% of male labor force)
SL.TLF.TERT.ZS	Labor force with tertiary education (% of total)
SL.TLF.TOTL.FE.ZS	Labor force, female (% of total labor force)
SL.TLF.TOTL.IN	Labor force, total
SL.UEM.1524.FE.ZS	Unemployment, youth female (% of female labor force ages 15-24)
SL.UEM.1524.MA.ZS	Unemployment, youth male (% of male labor force ages 15-24)
SL.UEM.1524.ZS	Unemployment, youth total (% of total labor force ages 15-24)
SL.UEM.LTRM.FE.ZS	Long-term unemployment, female (% of female unemployment)
SL.UEM.LTRM.MA.ZS	Long-term unemployment, male (% of male unemployment)
SL.UEM.LTRM.ZS	Long-term unemployment (% of total unemployment)
SL.UEM.PRIM.FE.ZS	Unemployment with primary education, female (% of female unemployment)
SL.UEM.PRIM.MA.ZS	Unemployment with primary education, male (% of male unemployment)
SL.UEM.PRIM.ZS	Unemployment with primary education (% of total unemployment)
SL.UEM.SECO.FE.ZS	Unemployment with secondary education, female (% of female unemployment)
SL.UEM.SECO.MA.ZS	Unemployment with secondary education, male (% of male unemployment)
SL.UEM.SECO.ZS	Unemployment with secondary education (% of total unemployment)
SL.UEM.TERT.FE.ZS	Unemployment with tertiary education, female (% of female unemployment)
SL.UEM.TERT.MA.ZS	Unemployment with tertiary education, male (% of male unemployment)
SL.UEM.TERT.ZS	Unemployment with tertiary education (% of total unemployment)
SL.UEM.TOTL.FE.ZS	Unemployment, female (% of female labor force)
SL.UEM.TOTL.MA.ZS	Unemployment, male (% of male labor force)
SL.UEM.TOTL.ZS	Unemployment, total (% of total labor force)
8. Urban development – 6 indicators.	

Cities can be tremendously efficient. It is easier to provide water and sanitation to people living closer together, while access to health, education, and other social and cultural services is also much more readily available. However, as cities grow, the cost of meeting basic needs increases, as does the strain on the environment and natural resources. Data on urbanization, traffic and congestion, and air pollution are from the United Nations Population Division, World Health Organization, International Road Federation, World Resources Institute, and other sources.	
INDICATOR_CODE	INDICATOR_NAME
EN.ATM.PM10.MC.M3	PM10, country level (micrograms per cubic meter)
EN.POP.DNST	Population density (people per sq. km of land area)
EN.URB.LCTY	Population in largest city
EN.URB.LCTY.UR.ZS	Population in the largest city (% of urban population)
EN.URB.MCTY	Population in urban agglomerations of more than 1 million
EN.URB.MCTY.TL.ZS	Population in urban agglomerations of more than 1 million (% of total population)
EP.PMP.DESL.CD	Pump price for diesel fuel (US$ per liter)
EP.PMP.SGAS.CD	Pump price for gasoline (US$ per liter)
IS.ROD.DESL.PC	Road sector diesel fuel consumption per capita (kt of oil equivalent)
IS.ROD.ENGY.ZS	Road sector energy consumption (% of total energy consumption)
IS.ROD.SGAS.PC	Road sector gasoline fuel consumption per capita (kt of oil equivalent)
IS.VEH.NVEH.P3	Motor vehicles (per 1,000 people)
IS.VEH.PCAR.P3	Passenger cars (per 1,000 people)
IS.VEH.ROAD.K1	Vehicles (per km of road)
SH.H2O.SAFE.UR.ZS	Improved water source, urban (% of urban population with access)
SH.STA.ACSN.UR	Improved sanitation facilities, urban (% of urban population with access)
SI.POV.URHC	Poverty headcount ratio at urban poverty line (% of urban population)
SP.URB.GROW	Urban population growth (annual %)
SP.URB.TOTL	Urban population
SP.URB.TOTL.IN.ZS	Urban population (% of total)

13 Appendix 2

The full dataset of 290 variables.

Dataset	Variable Name	Variable Label	Source	Stiglitz-Sen-Fitoussi area*
DSET1.PERSONAL_ACTIVITIES	FPR_15N19	Labur force participation rate - female 15-19 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	FPR_25N29	Labur force participation rate - female 25-29 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	FPR_55N59	Labur force participation rate - female 55-59 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	FPR_65P	Labur force participation rate - female 65+ years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	MFPR_25N29	Labur force participation rate - male/female 25-29 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	MFPR_30N34	Labur force participation rate - male/female 30-34 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	MFPR_35N39	Labur force participation rate - male/female 35-39 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	MFPR_40N44	Labur force participation rate - male/female 40-44 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	MFPR_45N49	Labur force participation rate - male/female 45-49 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	MFPR_50N54	Labur force participation rate - male/female 50-54 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	MFPR_55N59	Labur force participation rate - male/female 55-59 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	MPR_20N24	Labur force participation rate - male 20-24 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	MPR_30N34	Labur force participation rate - male 30-34 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	MPR_50N54	Labur force participation rate - male 50-54 years	International Labor Organization	4. Personal activities including work
DSET1.PERSONAL_ACTIVITIES	MPR_65P	Labur force participation rate - male 65+ years	International Labor Organization	4. Personal activities including work
DSET1.POLITICAL_VOICE	autoc	Institutionalized autocracy 0=low:10=high	Quality of Government Institute: Polity IV	5. Political voice and governance
DSET1.POLITICAL_VOICE	chga_demo	Democracy 0/1	Quality of Government Institute: Cheibub, Gandhi & Vreeland	5. Political voice and governance
DSET1.POLITICAL_VOICE	chga_hinst	Regime Institutions 0=parliamentary democracy 5=royal dictatorship	Quality of Government Institute: Cheibub, Gandhi & Vreeland	5. Political voice and governance
DSET1.POLITICAL_VOICE	democ	Institutionalized democracy	Quality of Government Institute Polity IV	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_auton	Autonomous regions 0/1	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_cemo	Chief Executive a military officer 0/1	Quality of Government Institute: Database of political Institutions	5. Political voice and governance

DSET1.POLITICAL_VOICE	dpi_checks	Number of Veto players	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_eipc	Executive index of political competitiveness	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_exelec	Executive election this year 0/1	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_finter	Finite term in office for chief executive 0/1	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_gps1	Largest government party number of seats	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_gps2	Second largest government party number of seats	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_gps3	Third largest government party number of seats	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_gs	Number of government seats	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_gvs	Government vote share (%)	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_legelec	Legislative election 0/1	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_lipc	Legislative index of political competitiveness	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_nogps	Number of other government party seats	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_noops	Number of other opposition party seats	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_nos	Number of oppositional seats	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_numul	Number of seats non-alligned / allegiance unknown	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_ovs	Opposition vote share (%)	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_seats	Total seats in legislature	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_slop1	Largest opposition party number of seats	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_slop2	Second largest opposition party number of seats	Quality of Government Institute: Database of political Institutions	5. Political voice and governance

DSET1.POLITICAL_VOICE	dpi_slop3	Third largest opposition party number of seats	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_system	Regime Type 0 direct presidential 1 strong president 2 parliamentary	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	dpi_yio	Number of years in office for Chief Executive	Quality of Government Institute: Database of political Institutions	5. Political voice and governance
DSET1.POLITICAL_VOICE	durable	Regime durability	Polity IV	5. Political voice and governance
DSET1.POLITICAL_VOICE	fh_cl	Civil Liberties 1 most free to 7 least free	Quality of Government Institute: Freedom House	5. Political voice and governance
DSET1.POLITICAL_VOICE	fh_ipolity2	Freedom House imputed polity 1- least democratic to 10 most decomatric	Quality of Government Institute: Freedom House	5. Political voice and governance
DSET1.POLITICAL_VOICE	fh_pr	Political Rights 1 most free 7 least free	Quality of Government Institute: Freedom House	5. Political voice and governance
DSET1.POLITICAL_VOICE	fh_status	Status 1 completely free 3 not free	Quality of Government Institute: Freedom House	5. Political voice and governance
DSET1.POLITICAL_VOICE	fragment	Polity fragmentation 0 no overt fragmentation to 3 serious fragmentation	Polity IV	5. Political voice and governance
DSET1.POLITICAL_VOICE	gd_ptss	Political terror scale, human rights score 1 to 5(most terror)	Quality of Government Institute: Gibney, Cornett & Wood - Political Terror Scale	5. Political voice and governance
DSET1.POLITICAL_VOICE	h_l1	1 if there is an effective legislative chamber else 0	Quality of Government Institute:Henisz - The political constraints index	5. Political voice and governance
DSET1.POLITICAL_VOICE	h_l2	1 if there is an effective legislative chamber else 0	Quality of Government Institute:Henisz - The political constraints index	5. Political voice and governance
DSET1.POLITICAL_VOICE	h_polcon3	Political constraints index 3: measures the feasibility of policy change	Quality of Government Institute:Henisz - The political constraints index	5. Political voice and governance
DSET1.POLITICAL_VOICE	ht_colonial	Colonial origin: 0 to 10	Quality of Government Institute: Hadenius & Teorell - Region and colonial origin	5. Political voice and governance
DSET1.POLITICAL_VOICE	ht_region	Region of the country 1 to 10	Quality of Government Institute: Hadenius & Teorell - Region and colonial origin	5. Political voice and governance
DSET1.POLITICAL_VOICE	ht_region2	Alternative Region of the Country 1 to 10 for contested cases	Quality of Government Institute: Hadenius & Teorell - Region and colonial origin	5. Political voice and governance
DSET1.POLITICAL_VOICE	ht_regspec	Regime Type (separating dominant multiparty systems)	Quality of Government Institute: Hadenius & Teorell - Region and colonial origin	5. Political voice and governance
DSET1.POLITICAL_VOICE	ht_regtype	Regime Type	Quality of Government Institute: Hadenius & Teorell - Region and colonial origin	5. Political voice and governance
DSET1.POLITICAL_VOICE	ht_regtype1	Regime Type collapsed	Quality of Government Institute: Hadenius & Teorell - Region and colonial origin	5. Political voice and governance
DSET1.POLITICAL_VOICE	pr	Political rights	Freedom House	5. Political voice and governance

DSET1.POLITICAL_VOICE	status	Freedom in the world	Freedom House	5. Political voice and governance
DSET1.SOCIAL_CONNECT	affected	Total Affected by Disaster	EM Disaster Database	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	al_ethnic	Ethnic fractionalization (probably that two randomly selected people belong two same ethnic group)	Quality of Government Institute: Alesina, Devleeschauwer, Easterly, Kurlat & Wacziarg	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	al_language	Language fractionalization (probably that two randomly selected people speak the same language)	Quality of Government Institute: Alesina, Devleeschauwer, Easterly, Kurlat & Wacziarg	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	al_religion	Religious fractionalization (probably that two randomly selected people belong to the same religion)	Quality of Government Institute: Alesina, Devleeschauwer, Easterly, Kurlat & Wacziarg	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	fe_cultdiv	Cultural diversity	Quality of Government Institute: Fearon	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	fe_etfra	Ethnic fractionalization (probably that two randomly selected people belong two same ethnic group)	Quality of Government Institute: Fearon	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	fe_plural	Population share of the largest ethnic group	Quality of Government Institute: Fearon	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	killed	Total killed by disaster	EM Disaster Database	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	lp_catho80	Catholics as percentage of the population in 1980	Quality of Government Institute: La Porta, Lopez-De-Silanes, Shleifer & Vishny	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	lp_lat_abst	Latitude	Quality of Government Institute: La Porta, Lopez-De-Silanes, Shleifer & Vishny	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	lp_legor	Legal Origin	Quality of Government Institute: La Porta, Lopez-De-Silanes, Shleifer & Vishny	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	lp_muslim80	Muslims as percentage of the population in 1980	Quality of Government Institute: La Porta, Lopez-De-Silanes, Shleifer & Vishny	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	lp_no_cpm80	Other denominations as percentage of the population 1980	Quality of Government Institute: La Porta, Lopez-De-Silanes, Shleifer & Vishny	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	lp_protmg80	Protestants as percentage of the population 1980	Quality of Government Institute: La Porta, Lopez-De-Silanes, Shleifer & Vishny	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	r_elf85	Ethno-linguistic fractionalization 1985	Quality of Government Institute: Roeder	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	wdi_area	Area (sq km)	Quality of Government Institute: World Development Indicators	6. Social connections and relationships
DSET1.SOCIAL_CONNECT	wdi_fr	Fertility rate (births per woman)	Quality of Government Institute: World Development Indicators	6. Social connections and relationships
DSET1.SECURITY	FPR_15N19	Labur force participation rate - female 15-19 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	FPR_25N29	Labur force participation rate - female 25-29 years	International Labor Organization	8. Insecurity

DSET1.SECURITY	FPR_55N59	Labur force participation rate - female 55-59 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	FPR_65P	Labur force participation rate - female 65+ years	International Labor Organization	8. Insecurity
DSET1.SECURITY	MFPR_25N29	Labur force participation rate - male/female 25-29 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	MFPR_30N34	Labur force participation rate - male/female 30-34 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	MFPR_35N39	Labur force participation rate - male/female 35-39 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	MFPR_40N44	Labur force participation rate - male/female 40-44 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	MFPR_45N49	Labur force participation rate - male/female 45-49 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	MFPR_50N54	Labur force participation rate - male/female 50-54 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	MFPR_55N59	Labur force participation rate - male/female 55-59 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	MPR_20N24	Labur force participation rate - male 20-24 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	MPR_30N34	Labur force participation rate - male 30-34 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	MPR_50N54	Labur force participation rate - male 50-54 years	International Labor Organization	8. Insecurity
DSET1.SECURITY	MPR_65P	Labur force participation rate - male 65+ years	International Labor Organization	8. Insecurity
DSET1.SECURITY	affected	Total Affected by Disaster	EM Disaster Database	8. Insecurity
DSET1.SECURITY	killed	Total Killed by Disaster	EM Disaster Database	8. Insecurity
DSET1.INEQUALITY	FPR_15N19	Labur force participation rate - female 15-19 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	FPR_25N29	Labur force participation rate - female 25-29 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	FPR_55N59	Labur force participation rate - female 55-59 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	FPR_65P	Labur force participation rate - female 65+ years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	MFPR_25N29	Labur force participation rate - male/female 25-29 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	MFPR_30N34	Labur force participation rate - male/female 30-34 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	MFPR_35N39	Labur force participation rate - male/female 35-39 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	MFPR_40N44	Labur force participation rate - male/female 40-44 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	MFPR_45N49	Labur force participation rate - male/female 45-49 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	MFPR_50N54	Labur force participation rate - male/female 50-54 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	MFPR_55N59	Labur force participation rate - male/female 55-59 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	MPR_20N24	Labur force participation rate - male 20-24 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	MPR_30N34	Labur force participation rate - male 30-34 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	MPR_50N54	Labur force participation rate - male 50-54 years	International Labor Organization	9. Inequality
DSET1.INEQUALITY	MPR_65P	Labur force participation rate - male 65+ years	International Labor Organization	9. Inequality
DSET1.EDUCATION	F15_No_Schooling	Percentage of female population 15+ who have no schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F15_Pop_N_000s	Number of female population 15+	Barro-Lee database of Educational Attainment in the World	3. Education

DSET1.EDUCATION	F15_Prim_Comp	Percentage of female population 15+ who have completed primary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F15_Prim_Tot	Percentage of female population 15+ whose highest level of education is primary	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F15_Sec_Com	Percentage of female population 15+ who have completed secondary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F15_Sec_Tot	Percentage of female population 15+ whose highest level of education is secondary	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F15_Tert_Com	Percentage of female population 15+ who have completed tertiary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F15_Tert_Tot	Percentage of female population 15+ whose highest level of education is tertiary	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F15_Year_Prim_School	Average number of years primary schooling for females 15+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F15_Year_Sec_School	Average number of years secondary schooling for females 15+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F15_Year_Tert_School	Average number of years tertiary schooling for females 15+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F15_Year_Tot_School	Average number of years total schooling for females 15+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F25_No_Schooling	Percentage of female population 25+ who have no schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F25_Pop_N_000s	Number of female population 25+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F25_Prim_Comp	Percentage of female population 25+ who have completed primary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F25_Prim_Tot	Percentage of female population 25+ whose highest level of education is primary	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F25_Sec_Com	Percentage of female population 25+ who have completed secondary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F25_Sec_Tot	Percentage of female population 25+ whose highest level of education is secondary	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F25_Tert_Com	Percentage of female population 25+ who have completed tertiary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F25_Tert_Tot	Percentage of female population 25+ whose highest level of education is tertiary	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F25_Year_Prim_School	Average number of years primary schooling for females 25+	Barro-Lee database of Educational Attainment in the World	3. Education

DSET1.EDUCATION	F25_Year_Sec_School	Average number of years secondary schooling for females 25+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F25_Year_Tert_School	Average number of years tertiary schooling for females 25+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	F25_Year_Tot_School	Average number of years total schooling for females 25+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_No_Schooling	Percentage of population 15+ who have no schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_Pop_N_000s	Number of population 15+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_Prim_Comp	Percentage of population 15+ who have completed primary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_Prim_Tot	Percentage of population 15+ whose highest level of education is primary	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_Sec_Com	Percentage of population 15+ who have completed secondary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_Sec_Tot	Percentage of population 15+ whose highest level of education is secondary	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_Tert_Com	Percentage of population 15+ who have completed tertiary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_Tert_Tot	Percentage of population 15+ whose highest level of education is tertiary	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_Year_Prim_School	Average number of years primary schooling for population 15+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_Year_Sec_School	Average number of years secondary schooling for population 15+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_Year_Tert_School	Average number of years tertiary schooling for population 15+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF15_Year_Tot_School	Average number of years total schooling for population 15+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF25_No_Schooling	Percentage of population 25+ who have no schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF25_Pop_N_000s	Number of population 25+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF25_Prim_Comp	Percentage of population 25+ who have completed primary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF25_Prim_Tot	Percentage of population 25+ whose highest level of education is primary	Barro-Lee database of Educational Attainment in the World	3. Education

DSET1.EDUCATION	MF25_Sec_Com	Percentage of population 25+ who have completed secondary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF25_Sec_Tot	Percentage of population 25+ whose highest level of education is secondary	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF25_Tert_Com	Percentage of population 25+ who have completed tertiary schooling	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF25_Tert_Tot	Percentage of population 25+ whose highest level of education is tertiary	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF25_Year_Prim_School	Average number of years primary schooling for population 25+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF25_Year_Sec_School	Average number of years secondary schooling for population 25+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	MF25_Year_Tert_School	Average number of years tertiary schooling for population 25+	Barro-Lee database of Educational Attainment in the World	3. Education
DSET1.EDUCATION	ihme_ayef	Average years of education (female)	Quality of Government Institute: Institute for health metrics and evaluation - University of Washington	3. Education
DSET1.EDUCATION	ihme_ayem	Average years of education (male)	Quality of Government Institute: Institute for health metrics and evaluation - University of Washington	3. Education
DSET1.ENVIRONMENT	_eco__agri_area_	Agricultural area square kilometres	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_eco__land_arabl	Arable land - square kilometres	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_eco__terr_prote	Protected areas - square kilometres	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__exchange_r	General exchange rate: local currency units per $US	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__fertility_	General fertility rate: average number of children a cohort of women could expect to have at the end of their reproductive life	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__infant_mor	Genereal infant mortality rate: number of deaths per thousand births of infants aged 0 to 1	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__land_area_	Land Area: Total area of the country excluding land under bodies of water	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__life_expec	General Life expectancy: The average number of years of life expectancy	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__migrants_n	Net Migrants: Net number of migrants, that is, the number of immigrants minus the number of emigrants (000s)	United Nations Environment Programme	7. Environment

DSET1.ENVIRONMENT	_gen__migration_	Net Migration rate: The number of immigrants minus the number of emigrants over a period, divided by the person-years lived by the population of the receiving country over that period. It is expressed as net number of migrants per 1,000 population.	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__mobile_pho	Number of phones per 100 inhabitants	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__pop_female	Defacto population as of 1 July of year indicated - female	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__pop_growth	Total increase of a population during a given period	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__pop_rural_	Population residing in rural areas 000's	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__pop_total_	Population: de facto population in a country, area or region as of 1 July of the year indicated (000's)	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_gen__pop_urban_	Population residing in urban areas 000's	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_res__agri_prod_	The FAO indices of agricultural production show the relative level of the aggregate volume of agricultural production for each year in comparison with the base period 1999-2001.	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_res__cereals_ha	Actual cereals harvested sq kilometres	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_res__cereals_pr	Cereal production - metric tons - includes grains and buckwheat	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_res__cereals_yi	Actual cereals yielded hectograms per hectare	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_res__eq_area_ir	Area equipped to provide water to the crops	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_res__fish_catch	Indice of relative level of the aggregate volume of fish caught for each year in comparison with the base period 1999-2001	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	_res__live_prod_	Indice of relative level of the aggregate volume of agricultural production for each year in comparison with the base period 1999-2001	United Nations Environment Programme	7. Environment
DSET1.ENVIRONMENT	affected	Total Affected by Disaster	EM Disaster Database	7. Environment
DSET1.ENVIRONMENT	killed	Total Killed by Disaster	EM Disaster Database	7. Environment
DSET1.AIDEFFECT2	DT_NFL_IFAD_CD	Net official flows from UN agencies, IFAD (current US$)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	DT_NFL_UNCF_CD	Net official flows from UN agencies, UNICEF (current US$)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	DT_NFL_UNCR_CD	Net official flows from UN agencies, UNHCR (current US$)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	DT_NFL_UNDP_CD	Net official flows from UN agencies, UNDP (current US$)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	DT_NFL_UNFP_CD	Net official flows from UN agencies, UNFPA (current US$)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	DT_NFL_UNTA_CD	Net official flows from UN agencies, UNTA (current US$)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	DT_NFL_WFPG_CD	Net official flows from UN agencies, WFP (current US$)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	DT_ODA_ALLD_KD	Net official development assistance and official aid received (constant 2008 US$)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	DT_ODA_ODAT_CD	Net official development assistance received (current US$)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	DT_ODA_ODAT_GI_ZS	Net ODA received (% of gross capital formation)	World Bank	1. Material living standards

DSET1.AIDEFFECT2	DT_ODA_ODAT_GN_ZS	Net ODA received (% of GNI)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	DT_ODA_ODAT_MP_ZS	Net ODA received (% of imports of goods and services)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	DT_ODA_ODAT_PC_ZS	Net ODA received per capita (current US$)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	EN_ATM_CO2E_PC	CO2 emissions (metric tons per capita)	World Bank	1. Material living standards
DSET1.AIDEFFECT2	IT_CEL_SETS_P2	Mobile cellular subscriptions (per 100 people)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NE_DAB_TOTL_CN	Gross national expenditure (current LCU)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NE_EXP_GNFS_CN	Exports of goods and services (current LCU)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NE_EXP_GNFS_ZS	Exports of goods and services (% of GDP)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NE_IMP_GNFS_ZS	Imports of goods and services (% of GDP)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NE_RSB_GNFS_CN	External balance on goods and services (current LCU)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NE_RSB_GNFS_ZS	External balance on goods and services (% of GDP)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NE_TRD_GNFS_ZS	Trade (% of GDP)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_ADJ_DCO2_CD	Adjusted savings: carbon dioxide damage (current US$)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_ADJ_DCO2_GN_ZS	Adjusted savings: carbon dioxide damage (% of GNI)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_ADJ_DKAP_CD	Adjusted savings: consumption of fixed capital (current US$)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_ADJ_DKAP_GN_ZS	Adjusted savings: consumption of fixed capital (% of GNI)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_ADJ_DMIN_CD	Adjusted savings: mineral depletion (current US$)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_ADJ_DMIN_GN_ZS	Adjusted savings: mineral depletion (% of GNI)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_ADJ_DNGY_CD	Adjusted savings: energy depletion (current US$)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_ADJ_DNGY_GN_ZS	Adjusted savings: energy depletion (% of GNI)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_GDP_DEFL_KD_ZG	Inflation, GDP deflator (annual %)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_GDP_DEFL_ZS	GDP deflator (base year varies by country)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_GDP_MKTP_CN	GDP (current LCU)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_GDP_MKTP_KD	GDP (constant 2000 US$)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_GDP_MKTP_KD_ZG	GDP growth (annual %)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_GDP_MKTP_KN	GDP (constant LCU)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_GDP_MKTP_PP_KD	GDP, PPP (constant 2005 international $)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_GDP_PCAP_CD	GDP per capita (current US$)	World Bank	1. Material living standards
DSET1.ECONOMICPOLICY3	NY_GDP_PCAP_KD	GDP per capita (constant 2000 US$)	World Bank	1. Material living standards
DSET1.EDUCATION4	SE_ENR_PRIM_FM_ZS	Ratio of female to male primary enrollment (%)	World Bank	3. Education
DSET1.EDUCATION4	SE_PRE_ENRR	School enrollment, preprimary (% gross)	World Bank	3. Education
DSET1.EDUCATION4	SE_PRM_AGES	Primary school starting age (years)	World Bank	3. Education
DSET1.EDUCATION4	SE_PRM_DURS	Primary education, duration (years)	World Bank	3. Education

DSET1.EDUCATION4	SE_PRM_ENRL	Primary education, pupils	World Bank	3. Education
DSET1.EDUCATION4	SE_PRM_ENRR	School enrollment, primary (% gross)	World Bank	3. Education
DSET1.ENVIRONMENT6	AG_PRD_CREL_MT	Cereal production (metric tons)	World Bank	7. Environment
DSET1.ENVIRONMENT6	EG_ELC_FOSL_ZS	Electricity production from oil, gas and coal sources (% of total)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_ADJ_AEDU_CD	Adjusted savings: education expenditure (current US$)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_ADJ_AEDU_GN_ZS	Adjusted savings: education expenditure (% of GNI)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_ADJ_DCO2_CD	Adjusted savings: carbon dioxide damage (current US$)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_ADJ_DCO2_GN_ZS	Adjusted savings: carbon dioxide damage (% of GNI)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_ADJ_DFOR_CD	Adjusted savings: net forest depletion (current US$)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_ADJ_DFOR_GN_ZS	Adjusted savings: net forest depletion (% of GNI)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_ADJ_DKAP_GN_ZS	Adjusted savings: consumption of fixed capital (% of GNI)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_ADJ_DMIN_CD	Adjusted savings: mineral depletion (current US$)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_ADJ_DMIN_GN_ZS	Adjusted savings: mineral depletion (% of GNI)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_ADJ_DNGY_CD	Adjusted savings: energy depletion (current US$)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_ADJ_DNGY_GN_ZS	Adjusted savings: energy depletion (% of GNI)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_GDP_COAL_RT_ZS	Coal rents (% of GDP)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_GDP_FRST_RT_ZS	Forest rents (% of GDP)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_GDP_MINR_RT_ZS	Mineral rents (% of GDP)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_GDP_NGAS_RT_ZS	Natural gas rents (% of GDP)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_GDP_PETR_RT_ZS	Oil rents (% of GDP)	World Bank	7. Environment
DSET1.ENVIRONMENT6	NY_GDP_TOTL_RT_ZS	Total natural resources rents (% of GDP)	World Bank	7. Environment
DSET1.FINANCIALSECTOR7	BX_KLT_DINV_CD_WD	Foreign direct investment, net inflows (BoP, current US$)	World Bank	1. Material living standards
DSET1.FINANCIALSECTOR7	FI_RES_TOTL_CD	Total reserves (includes gold, current US$)	World Bank	1. Material living standards
DSET1.FINANCIALSECTOR7	FM_AST_DOMS_CN	Net domestic credit (current LCU)	World Bank	1. Material living standards
DSET1.FINANCIALSECTOR7	FM_AST_NFRG_CN	Net foreign assets (current LCU)	World Bank	1. Material living standards
DSET1.FINANCIALSECTOR7	PA_NUS_ATLS	DEC alternative conversion factor (LCU per US$)	World Bank	1. Material living standards
DSET1.FINANCIALSECTOR7	PA_NUS_FCRF	Official exchange rate (LCU per US$, period average)	World Bank	1. Material living standards
DSET1.LABOUR10	SL_TLF_CACT_FE_ZS	Labor participation rate, female (% of female population ages 15+)	World Bank	4. Personal activities including work
DSET1.LABOUR10	SL_TLF_CACT_MA_ZS	Labor participation rate, male (% of male population ages 15+)	World Bank	4. Personal activities including work
DSET1.LABOUR10	SL_TLF_CACT_ZS	Labor participation rate, total (% of total population ages 15+)	World Bank	4. Personal activities including work
DSET1.LABOUR10	SL_TLF_TOTL_FE_ZS	Labor force, female (% of total labor force)	World Bank	4. Personal activities including work
DSET1.LABOUR10	SL_TLF_TOTL_IN	Labor force, total	World Bank	4. Personal activities including work
DSET1.URBANDEVELOP16	EN_POP_DNST	Population density (people per sq. km of land area)	World Bank	7. Environment

DSET1.URBANDEVELOP16	EN_URB_LCTY	Population in largest city	World Bank	7. Environment
DSET1.URBANDEVELOP16	EN_URB_LCTY_UR_ZS	Population in the largest city (% of urban population)	World Bank	7. Environment
DSET1.URBANDEVELOP16	SP_URB_GROW	Urban population growth (annual %)	World Bank	7. Environment
DSET1.URBANDEVELOP16	SP_URB_TOTL	Urban population	World Bank	7. Environment
DSET1.URBANDEVELOP16	SP_URB_TOTL_IN_ZS	Urban population (% of total)	World Bank	7. Environment
DSET1.FULLHEALTH	SH_DYN_AIDS_ZS	Prevalence of HIV, total (% of population ages 15-49)	World Bank	2. Health
DSET1.FULLHEALTH	SH_IMM_IDPT	Immunization, DPT (% of children ages 12-23 months)	World Bank	2. Health
DSET1.FULLHEALTH	SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)	World Bank	2. Health
DSET1.FULLHEALTH	SH_TBS_DTEC_ZS	Tuberculosis case detection rate (%, all forms)	World Bank	2. Health
DSET1.FULLHEALTH	SH_TBS_INCD	Incidence of tuberculosis (per 100,000 people)	World Bank	2. Health
DSET1.FULLHEALTH	SH_XPD_EXTR_ZS	External resources for health (% of total expenditure on health)	World Bank	2. Health
DSET1.FULLHEALTH	SH_XPD_OOPC_TO_ZS	Out-of-pocket health expenditure (% of total expenditure on health)	World Bank	2. Health
DSET1.FULLHEALTH	SH_XPD_OOPC_ZS	Out-of-pocket health expenditure (% of private expenditure on health)	World Bank	2. Health
DSET1.FULLHEALTH	SH_XPD_PCAP_PP_KD	Health expenditure per capita, PPP (constant 2005 international $)	World Bank	2. Health
DSET1.FULLHEALTH	SH_XPD_PRIV_ZS	Health expenditure, private (% of GDP)	World Bank	2. Health
DSET1.FULLHEALTH	SH_XPD_PUBL	Health expenditure, public (% of total health expenditure)	World Bank	2. Health
DSET1.FULLHEALTH	SH_XPD_PUBL_GX_ZS	Health expenditure, public (% of government expenditure)	World Bank	2. Health
DSET1.FULLHEALTH	SH_XPD_PUBL_ZS	Health expenditure, public (% of GDP)	World Bank	2. Health
DSET1.FULLHEALTH	SH_XPD_TOTL_ZS	Health expenditure, total (% of GDP)	World Bank	2. Health
DSET1.FULLHEALTH	SP_DYN_CBRT_IN	Birth rate, crude (per 1,000 people)	World Bank	2. Health
DSET1.FULLHEALTH	SP_DYN_CDRT_IN	Death rate, crude (per 1,000 people)	World Bank	2. Health
DSET1.FULLHEALTH	SP_DYN_LE00_FE_IN	Life expectancy at birth, female (years)	World Bank	2. Health
DSET1.FULLHEALTH	SP_DYN_LE00_MA_IN	Life expectancy at birth, male (years)	World Bank	2. Health
DSET1.FULLHEALTH	SP_DYN_TFRT_IN	Fertility rate, total (births per woman)	World Bank	2. Health
DSET1.FULLHEALTH	SP_POP_0014_TO_ZS	Population ages 0-14 (% of total)	World Bank	2. Health
DSET1.FULLHEALTH	SP_POP_1564_TO_ZS	Population ages 15-64 (% of total)	World Bank	2. Health
DSET1.FULLHEALTH	SP_POP_65UP_TO_ZS	Population ages 65 and above (% of total)	World Bank	2. Health
DSET1.FULLHEALTH	SP_POP_GROW	Population growth (annual %)	World Bank	2. Health
DSET1.FULLHEALTH	SP_POP_TOTL	Population, total	World Bank	2. Health
DSET1.FULLHEALTH	SP_POP_TOTL_FE_ZS	Population, female (% of total)	World Bank	2. Health

*NB: Indicators can be processed within multiple areas.

14 Appendix 3

Included in this appendix is an example SAS program showing the genetic algorithm / spectral clustering implementation from section 7.2, which may contain code that is of use to other researchers. Comments explaining the process are shown in green within the code.

```
************************************************************;
*Filename:   Spectral Clustering GA with hist 1996.sas                    ;
*Author:     Lisa Henley                                                  ;
*Date:       7 February 2013                                              ;
*Description: To run a spectral clustering solution on the               ;
*       flourishing dataset                                               ;
************************************************************;
*suppress log once program has run error free;
options nonotes nosource nosource2 errors=0 ;
*options notes source source2 errors=0 ;

%let startime  = %sysfunc(time(),time8.);
options nomprint nomlogic nosymbolgen spool;
*options mprint mlogic symbolgen spool;

*set libname;
libname dset1 "E:\lhe11\";

*set macro variables;
%let inv_year = 1996;
%let inv_year_last = %eval(&inv_year. - 1);
%let count_vars = 105; *number of variables;
%let allele = %eval(&count_vars.);
%let initpop=100;
%let reps = %eval(&initpop./2);
%let top_count = 40; *the amount selected for crossover;
%let elite = 2; *top options which are kept from generation to generation;
%let crossreps =%eval(( %eval( &top_count.) - %eval(&elite.)) / 2);
%let maxit =150; *maximum number of iterations;
%let stable = 20;  *number of 'same' iterations before solution judged stable;
%let restrict_vars = 30; *maximum number of variables;
%let restrict_last = 20; *number of variables that must match to last time;
%let restrict_kslast = 30; *number of variables in the canonical solution must match to last
time;
%let min_clus_size = 20;

*keep a copy of dset for profiling;
proc sort data= dset1.alltogethernow out=flourish_&inv_year.;
        where year = "_&inv_year.";
        by country;
```

```
run;


*get variable descriptions;
proc import
	datafile = "P:\Data\DatasetOne\outfiles\Distribution of variables in final
dataset.xls"
	dbms = xls out=vardescript replace;
quit;

proc sort data=vardescript (keep=_stat_ description)
	out=vardescript (rename=(_stat_ = _var_));
	by _stat_;
run;

*create the dataset;
data flourishingfull_&inv_year.;
	set dset1.alltogethernow;
	where year = "_&inv_year.";
run;

*bring in the descriptive vars from 1995 for use in analysis later;
proc import datafile="P:\Data\DatasetOne\outfiles\Transfer\fit_spectral.xlsx"
	dbms = xlsx out=fit_spectral95 replace;
	sheet="Vars with good separation" ;
quit;

*keep the best numerica and categoricals...these were hand picked;
*from the original box plot analysis;
data fit_spectral95 (keep=_var_);
	set fit_spectral95;
	where boxplot_best = "Keep" or
		cat_best = "KeepC";
	_var_ = upcase(_var_);
run;

*sort ready for a merge later;
proc sort data=fit_spectral95;
	by _var_;
run;

*clean up the workspace;
%macro cleanup(dname);
%if %sysfunc(exist(&dname)) %then %do;
   proc datasets library=work noprint;
		delete &dname;
	quit;
	run;
%end;
%mend;

%cleanup(elite);
%cleanup(nextgen);
%cleanup(oldies);
```

```
%cleanup(holding);
%cleanup(allfit);
%cleanup(winning_detail);

*convert the categorical variables to binary;
%macro convertbinary(varname);
*find the class levels;
proc freq data=flourishingfull_&inv_year. noprint;
	table &varname / out=classes;
run;

*store the class names and how many there are;
proc sql noprint;
select &varname into :var1 - :var&SysMaxLong from classes;
%let dimvar = &SqlObs;
quit;

*now make the changes in the dataset;
data flourishingfull_&inv_year. (drop=i &varname);
	set flourishingfull_&inv_year.;
	array A &varname.1 - &varname.&dimvar;
	do i = 1 to &dimvar.;
		A(i) = (&varname = i);
	end;
run;

*delete any categories that are empty or contain too little data;
%do i = 1 %to &dimvar;
	proc sql noprint;
	select sum(&varname&i) into :hascontent from flourishingfull_&inv_year.;
	quit;

	%if %eval(&hascontent) <25  %then %do;
	data flourishingfull_&inv_year. (drop = &varname&i);
		set flourishingfull_&inv_year.;
	run;
	%put dropped &varname&i. ;
	%end;
%end;

*change so records insert rather than creating a new table;
%let bin_count = 2;

%mend;

%convertbinary(ht_region);
%convertbinary(ht_regtype1);
%convertbinary(gd_ptss);
%convertbinary(fh_pr);
%convertbinary(fh_cl);
%convertbinary(dpi_system);
%convertbinary(dpi_numul);
%convertbinary(dpi_lipc);
%convertbinary(dpi_checks);
```

```
%convertbinary(dpi_cemo);
%convertbinary(chga_hinst);
%convertbinary(ht_colonial);
%convertbinary(lp_legor);

*get the counts of the variables;
proc contents data=flourishingfull_&inv_year.
        out=var_names(keep=name) noprint;
run;

**********needs manual intervention here for numbering;
* only keep the independent variable names;
data var_names ;
        informat _NAME_ $9.;
        set var_names;
        if name not in ('year','country');
        count+1;
        *need to work with categories now.  Follow "orderofvarnames.xls";
        if index(name, "chga_hinst") > 0 then count = 85;
        /*if index(name, "dpi_cemo") > 0 then count = 86;*/
        if index(name, "dpi_checks") > 0 then count = 86;
        if index(name, "dpi_lipc") > 0 then count = 87;
        /*if index(name, "dpi_numul") > 0 then count = 89;*/
        if index(name, "dpi_system") > 0 then count = 88;
        if index(name, "durable") > 0 then count = 89;
        if index(name, "fe_etfra") > 0 then count = 90;
        if index(name, "fe_plural") > 0 then count = 91;
        if index(name, "fh_cl") > 0 then count = 92;
        if index(name, "fh_pr") > 0 then count = 93;
        if index(name, "gd_ptss") > 0 then count = 94;
        if index(name, "ht_colonial") > 0 then count = 95;
        if index(name, "ht_region") > 0 then count = 96;
        if index(name, "ht_regtype1") > 0 then count = 97;
        if index(name, "killed") > 0 then count = 98;
        if index(name, "lp_catho80") > 0 then count = 99;
        if index(name, "lp_lat_abst") > 0 then count = 100;
        if index(name, "lp_legor") > 0 then count = 101;
        if index(name, "lp_muslim80") > 0 then count = 102;
        if index(name, "lp_no_cpm80") > 0 then count = 103;
        if index(name, "lp_protmg80") > 0 then count = 104;
        if index(name, "wdi_fr") > 0 then count = 105;

        *now create variable for merge;
        _NAME_ = compress("allele" || count);
run;

*sort for merge;
proc sort data=var_names; by _NAME_; run;

*Create affinity matrix for later use in spectral clustering. ;
*firstly calculate rowwise correlations (kendals tau and spearmans);
*then potentially do this between columns for bivariate graph;
proc sort data=flourishingfull_&inv_year.;
        by country;
```

```
run;

*keep country names;
data names (keep = country);
	set flourishingfull_&inv_year.;
run;

*make the variable names a consistent case;
data upcase_var_names;
	set var_names;
	NAME = upcase(name);
run;

*only keep names from last years winner;
proc sql noprint;
	create table last_years_record as
	select upcase_var_names.NAME, upcase_var_names.count from
	upcase_var_names as a inner join dset1.history&inv_year_last as b
	on a.NAME = b.model;

	select count into :lastvarlist separated by ", "
	from last_years_record;
quit;

**********needs manual intervention here**********;
**********some variables may have been dropped alltogether;

data initpop (drop = i j temp);
	array allele {&allele}  ;
	do until (_n_ = %eval(&initpop )) ;
		*first choose which variables will be in the model;
		do i = 1 to &count_vars; /*working just with variable selection
			there are about 135 countries so want to choose
			1/5th of the 10x variables....round down to say 25 of them*/
			temp=uniform(1);
			*select 15% to get meaningful correlations;
			if temp <0.1 then allele[i] = 1;
			else allele[i]=0;
		end;
		*weight the last years winning variables;
		do j = &lastvarlist ;
			if temp < 0.6 then allele[j] = 1;
		end;

		*make a chromosone of the alleles relating to ;
		*the eigenvector selection, last two not included;
		chromosone = catt(of allele1-allele&allele.);
		output;
		_n_+1;
	end;
run;

*add the previous year's record for selection;
*make holding table for all alleles;
```

```
data hold_alleles;
        do until (count = &count_vars);
        count+1;
        output;
        end;
run;

*sort for merge;
proc sort data=hold_alleles; by count; proc sort data=last_years_record nodupkey; by
count; run;
run;

data merge_last (keep=allele);
        merge hold_alleles (in=a) last_years_record (in=b);
        by count;
        if b then allele = 1;
        else allele = 0;
run;

proc transpose data=merge_last out=merge_last_trans prefix=allele;
run;

data merge_last_trans (drop=_NAME_);
        set merge_last_trans;
        chromosone = catt(of allele1-allele&allele.);
        replicate= &reps ;
run;

/*select only 20 variables
build dataset of samples.  The sample size is two because the
selection method is binary tournament, so each replicate has
two members*/

*this method of sampling select ALL of the top chromosones;
*to increase speed;
proc surveyselect data=initpop method = SRS rep = 2
                sampsize = &reps seed = 12345
                                        out = testfit noprint;
        id _all_;
run;

*number into replicates;
data testfit;
        set testfit;
        by replicate;
        if first.replicate then rep2 = 1;
        else rep2+1;
run;

*now re-sort;
proc sort data=testfit;
        by rep2;
run;
```

```
*make new number the replicate number;
data testfit (rename=(rep2 = replicate));
	set testfit (drop = replicate);
run;

*resort;
proc sort data=testfit;
	by replicate;
run;

*need to replace one test subject with last year's winner;
data testfit;
	set testfit end = eof;
	if eof then delete;
run;

*add last year's winner;
data testfit;
	set testfit merge_last_trans;
run;

*sort for numbering;
proc sort data=testfit;
	by replicate;
run;

/*create a variable in sample dataset which records which partner in
tournament each model is i.e. each sample (tournament) has two
competitors (models) and they are numbered 1 or 2*/
data testfit;
	set testfit;
	by replicate;
	if first.replicate then option=1;
	else option=2;
run;

*this is the tournament macro which is called later;
/*this macro runs the tournament*/
%macro tournie();
	%do i = 1 %to 2;
		proc sql noprint;
			select compress(opt&i.) into :opt&i._vars separated by " "
			from optionnames;
		quit;

		proc sort data=flourishingfull_&inv_year. out=flourish_opt&i
				(keep=country year &&opt&i._vars.);
			by country;
		run;

		*check there are some variables;
		%let dsid = %sysfunc(open(flourish_opt&i ));
		%let nvars=%sysfunc(attrn(&dsid,NVARS));
		%let rc = %sysfunc(close(&dsid));
```

```
*exit if there arent enough variables;
%if &nvars <3 %then %do;
	%let kdvalue = 0;
	%goto continue;
%end;

proc stdize data=flourish_opt&i. out=flourish_opt&i
	method = MAD reponly nomiss;
run;

*transpose so countries are in columns;
proc transpose data=flourish_opt&i. out=rowwise;
	id country;
	var _numeric_;
run;

*work out kendell correlations;
proc corr data= rowwise outk=row_kendall noprint;
	var _numeric_;
run;

*spectral clustering
*calculate affinity matrix for fully connected graph;
data affinity_country (drop=_type_  i);
	set row_kendall;
	if _type_='CORR' ;
	*change the 1s on the diagonals to 0s;
	array numlist(*) _numeric_;
	do i = 1 to dim(numlist);
	 if numlist(i) = 1 THEN numlist(i)=0;
	end;
run;

*the correlation is between -1 and 1 but we need weights that;
*are positive.  try adding 1 to all weights except 0;
data affinity_country_scaled (drop=i);
	set affinity_country;
	array numlist(*) _numeric_;
	do i = 1 to dim(numlist);
		if numlist(i) ne 0 then numlist(i) + 1;
	end;
run;

*create weighted degree matrix;
data distance_country(keep=_NAME_ weighted_dist);
	set affinity_country_scaled;
	weighted_dist = sum(of Albania--Zambia);
run;

proc iml;

	*diagonalise distance matrix - ;
	*create a normalised version;
```

```
        use distance_country ;
        read all  ;

        Ndeg=diag(1/weighted_dist);

        *get affinity matrix;
        use affinity_country_scaled;

        read all var _num_ into affinity_scaled;

        *create normalised laplacian;
        *Lsym := D^-1/2LD^-1/2 = I - D^-1/2W D^-1/2;
        *Lrw :=D^-1L=I-D^-1W (used below);
        NLaplacian = I(nrow(affinity_scaled)) - Ndeg*affinity_scaled;

        *get the eigenvectors and values;
        call eigen(values,vectors,NLaplacian);

        create Nvectors from vectors ;
        append from vectors;

quit;

*The last eigenvector has entries that are all zero as this is a ;
*fully connected graph.  Drop this last eigenvector and use the;
*optimal cut vector;
data nvectors2 (keep = col134 col133 col132);
        set nvectors ;
run;

*cluster using k-means;
proc fastclus data=Nvectors2 noprint
        out=clusters maxc=3 maxiter=50;
run;

*store solution;
data cluster ;
        merge names clusters (keep=cluster) ;
run;

*check counts are reasonably balanced;
proc freq data=cluster noprint ;
        table cluster / out=freq_check;
run;

proc sql noprint;
select min(percent) into :freq_check from freq_check;
quit;

%if %sysevalf(&freq_check) < %eval(&min_clus_size) %then %do;

data joined_fit&i;
        option = &i;
        last_fit = -999;
```

```
        curr_fit = -999;
        count_current = 999;
        ks_agree95 = -999;
run;

%goto endtrial;


%end;

*merge the cluster onto the original dataset for profiling;
data cluster_profile;
        merge flourishingfull_&inv_year. cluster;
        by country;
run;

proc npar1way edf data=cluster_profile
        median wilcoxon savage noprint;
   class cluster;
        output out=AllKSStat  median wilcoxon edf savage;
run;

*find how well the top 30 match last years top 30;
*need to preclude categoricals...they are selected if significant;
*at 0.001 level;
*first find overall total;
data allksstatchecklastyear;
        set allksstat;
        *initialise variables for categoricals;
        categorical = 0;
        sig_overall = 0;
        overall = sum(_KW_, _CHMED_, _CHSAV_, _KS_);
        *flag if variable is categorical;
        if index(_var_,"ht_region") > 0 or
        index(_var_,"ht_regtype1") > 0 or
        index(_var_,"gd_ptss") > 0 or
        index(_var_,"fh_pr") > 0 or
        index(_var_,"fh_cl") > 0 or
        index(_var_,"dpi_system") > 0 or
        index(_var_,"dpi_numul") > 0 or
        index(_var_,"dpi_lipc") > 0 or
        index(_var_,"dpi_checks") > 0 or
        index(_var_,"dpi_cemo") > 0 or
        index(_var_,"chga_hinst") > 0 or
        index(_var_,"ht_colonial") > 0 or
        index(_var_,"lp_legor") > 0 then categorical = 1;
        *now flag significant categoricals;
        if p_kw < 0.001 and p_chmed < 0.001 and p_chsav <0.001
                then sig_overall = 1;
run;

*get top numerics;
proc sql outobs=30;
create table top30current_numeric as
```

```
select _var_ from allksstatchecklastyear
where categorical = 0
order by overall desc;

*get top categorical;
create table topcurrent_categorical as
select _var_ from allksstatchecklastyear
where sig_overall = 1;

quit;

*whatare the top variables;
data topcurrents;
	set top30current_numeric topcurrent_categorical;
	_var_ = upcase(_var_);
run;

*sort for merge;
proc sort data=topcurrents;
by _var_;
run;

*merge this replicate ks stats with the 1995 stats;
data ks_agree_95;
	merge fit_spectral95 (in=a) topcurrents (in=b);
	by _var_;
	if a and b then output;
run;

*now do the actual KS analysis;
*summarise results keep the mean and the lowest quartile value;
*want to maximise both;
proc summary data=AllKSStat nway missing;
	var _KW_ _CHMED_ _CHSAV_ _KS_;
	output out = AllKSStat (drop=_type_ _freq_) mean= p25= /
autoname;
run;

*transpose for comparison;
proc transpose data=AllKSStat out=AllKSStat (rename=(COL1 = opt&i));
run;

proc sql noprint;
	select sum(opt&i) into :fitness&i from ALLKSStat;
	*also count number of ks good differentiators that matched;
	*to original year (1995);
	select count(_var_) into :ks_agree95 from ks_agree_95;
	select _var_ into :canonlist separated by " " from topcurrents;
quit;
*store the total value for each model in to macro variables;
*(want to maximise);
%let kdvalue = 0;

*check how well the model variables match the historical winner;
```

```
*get the current option names;
proc sort data=optionnames out=currentvars (rename=(opt&i = model));
	by opt&i;
	where opt&i ne "";
run;

*make name upper case;
data currentvars;
	set currentvars;
	model = upcase(model);
run;

*sort for merge;
proc sort data=currentvars;
	by model;
run;

*match them with the historical names;
data match (keep=model);
	merge currentvars (in=a) dset1.history&inv_year_last (in=b) ;
	by model;
	if a and b then output;
run;

*how many match;
proc sql noprint;
	select count(model) into :fitness_last&i from match;
	select model into :modellist separated by " " from match;
	select count(model) into :count_current from currentvars;
	select count(model) into :count_historical from
dset1.history&inv_year_last;
	%let dimvar = &SqlObs;
quit;

*add the current and historical fitness together;
data joined_fit&i;
	informat variables canonlist $1000.;
	format variables canonlist $1000.;
	variables = " &modellist ";
	last_fit = &&fitness_last&i;
	curr_fit = &&fitness&i;
	count_current = &count_current;
	ks_agree95 = &ks_agree95;
	canonlist = " &canonlist ";
	freq_check = &freq_check;
	option = &i;
	addee = 1; *flag to indicate if this is the best;
run;

*if the number of variables in the current solution is too high;
*and the number that match to history too low then;
*make the variate lose;
data joined_fit&i;
	set joined_fit&i;
```

```
        if count_current > %eval(&restrict_vars) or
                last_fit < %eval(&restrict_last) or
                ks_agree95 < %eval(&restrict_kslast) then do;
                last_fit = -999;
                curr_fit = -999;
                count_current = 999;
                ks_agree95 = -999;
        end;
run;


*reset the addee flag;
%if %sysfunc(exist(allfit)) %then %do;
data allfit;
        set allfit;
        addee = 0;
run;
%end;

proc append base = allfit data=joined_fit&i ;
quit;

*rank the variables for the fitness function;
proc rank data=allfit   out = allfit_ranked;
        var last_fit curr_fit count_current ks_agree95;
        ranks last_rank curr_rank count_rank kslast_rank;
run;

*calculate fitness;
data allfit_ranked;
        set allfit_ranked;
        *want to maximise the following;
        *the rank variables matched to the last round ;
        *(higher is better) the rank of the current fitness;
        *(higher is better) and subtract the rank of the number;
        *of variables (lower is better);
        fitness = last_rank  + curr_rank + kslast_rank - count_rank;
run;


*store the results of the analysis if this replicate has;
*the greatest fitness so far.  Restrict to one record;
*no point replacing result if it is the same fitness;
proc sql outobs=1 noprint;
create table isthisgreatest as
select addee from allfit_ranked
having fitness = max(fitness);

select addee into :check from isthisgreatest;

quit;

*now if the current record has greater fitness, store the;
*profile and history;
```

```
        %if %eval(&check) = 1 %then %do;
        data dset1.winning_cluster_profile_&inv_year;
                set cluster_profile;
        run;

        *keep the history for the evolutionary process;
        data dset1.history&inv_year (keep=model fitness);
                set optionnames;
                where opt&i ne " ";
                model=compress(upcase(Name));
                fitness = &&fitness&i;
        run;

        data dset1.winning_detail_&inv_year ;
                informat winning_vars canonlist $1000.;
                format winning_vars canonlist $1000.;
                winning_vars =" &&opt&i._vars. ";
                matched_vars = &&fitness_last&i ;
                count_current = &count_current;
                count_historical = &count_historical;
                ks_agree95 = &ks_agree95;
                canonlist = " &canonlist ";
                freq_check = &freq_check;
        run;



        %end;
%endtrial:
%end;

*find the winner out of the two;
data joined_fit;
        set joined_fit1 joined_fit2;
run;

proc rank data=joined_fit   out = joined_fit_ranked;
        var last_fit curr_fit count_current ks_agree95;
        ranks last_rank curr_rank count_rank kslast_rank;
run;

data joined_fit_ranked;
        set joined_fit_ranked;
        fitness = last_rank  + curr_rank + kslast_rank - count_rank;
run;

proc sort data=joined_fit_ranked;
        by descending fitness;
run;

data joined_fit_ranked;
        set joined_fit_ranked;
        if _n_ = 1;
run;
```

```
        proc sql noprint;
        select option into :win from joined_fit_ranked;
        select last_fit into :last_fit from joined_fit_ranked;
        select curr_fit into :curr_fit from joined_fit_ranked;
        select count_current into :count_curr from joined_fit_ranked;
        select ks_agree95 into :ks_agree95 from joined_fit_ranked;
        quit;

        data winner;
                option = &win.;
                replicate = &trial.;
                last_fit = &last_fit.;
                curr_fit = &curr_fit.;
                count_current = &count_curr.;
                ks_agree95 = &ks_agree95.;
        run;

        *if the number of variables in the current solution is too high;
        *make the variate lose;
        data winner;
                set winner;
                if count_current > %eval(&restrict_vars) or
                                last_fit < %eval(&restrict_last)or
                                ks_agree95 < %eval(&restrict_kslast) then do;
                        last_fit = -999;
                        curr_fit = -999;
                        count_current = 999;
                        ks_agree95 = -999;
                end;
        run;

        *store results;
        proc append data = winner base = win_results;
        run;
        %continue:

%mend;

*first run uses 50 loops;
%let loops = %eval(&reps);

*transpose sample for merge;
%macro fitness();
*for each of the tournaments;
%do trial = 1 %to &loops;

        *transpose the sample for ease of working;
        proc transpose data=testfit out=testfitturn;
                where replicate =  &trial;
        run;

        *sort for merge;
```

```
	proc sort data=testfitturn; by _NAME_; run;
	proc sort data=var_names; by _name_; run;

	/*only keep variable names where the allele (col1 or col2) is flagged 1
	store these kept names in variables 'opt1' and 'opt2'.  If an allele has
	value 1 then that variable will be in the model, if it is 0 then that
	variable won't be in the model*/
	data optionnames;
		merge testfitturn (where = (_name_ ne 'Replicate'
			and _name_ ne 'option' and _name_ ne 'chromosone')) var_names;
			by _NAME_;
		*variable names;
		if col1 = 1 then opt1 = name;
		if col2 = 1 then opt2 = name;
	run;

	/*compress these two variables into a string and store in a macro
	variable for use in next step
	initialise the standardisation variables for each round*/
	%tournie();



%end;


	proc rank data=win_results   out = win_results_ranked;
		var last_fit curr_fit count_current ks_agree95;
		ranks last_rank curr_rank count_rank kslast_rank;
	run;

	data win_results_ranked;
		set win_results_ranked;
		fitness = last_rank  + curr_rank + kslast_rank - count_rank;
	run;



%mend;
%fitness()



%macro crossmut();
	*now onto the next generation...merge the winners with their ;
	*respective alleles and take best 50percentish;
	*if this is not the first generation then take a copy of ;
	*nextgen as these are the parents;
	%if %sysfunc(exist(nextgen)) %then %do;
		data oldies;
			set nextgen;
		run;
	%end;
```

```
*if this is not the first run then put the elites back in the mix;
*the outobs is only necessary for the first run, after this;
*there will always be half this amount in this table;
*we are taking out the top performers;
proc sql outobs= &top_count noprint;
        create table nextgen as
        select chromosone, fitness
        from testfit inner join win_results_ranked
        on testfit.option = win_results_ranked.option and
                testfit.replicate = win_results_ranked.replicate
        order by 2 desc; /*want highest value*/
quit;

*add the former elites  into the mix;
%if %sysfunc(exist(elite)) %then %do;
data nextgen;
                set nextgen elite;
        run;
%end;

*and the former generations best candidates;
%if %sysfunc(exist(oldies)) %then %do;
data nextgen;
                set nextgen oldies;
        run;
%end;

*re-sort and keep best  - keep the new best as elite;
*if using that option;
proc sql outobs= &top_count noprint ;
        create table holding as /*use new name to stop warning*/
        select * from nextgen
        order by 2 desc; /*want highest value*/
quit;

*keep the new elites (if using option), this will overwrite the old elite;
data elite nextgen;
        set holding;
        if _n_ <= &elite then output elite;
        else output nextgen;
run;

*now create samples for crossover;
*this method of sampling select ALL of the top chromosones;
*to increase speed;
proc surveyselect data=nextgen method = SRS rep = 2
        sampsize = &crossreps seed = 12345
                                        out = crossoverg2 (keep = chromosone
replicate) noprint;
        id _all_;
run;

*number into replicates;
data crossoverg2;
```

```
        set crossoverg2;
        by replicate;
        if first.replicate then rep2 = 1;
        else rep2+1;
run;

*now re-sort;
proc sort data=crossoverg2;
        by rep2;
run;

*make new number the replicate number;
data crossoverg2 (rename=(rep2 = replicate));
        set crossoverg2 (drop = replicate);
run;

*resort;
proc sort data=crossoverg2;
        by replicate;
run;


*the code from this point to ***** is the crossover process;
*start the crossover process, create pieces to swap;
data crossoverg2 swappiece (keep = replicate option swap);
        retain crosspoint;
        set crossoverg2;
        by replicate;
        *make a random crossover point for each replicate group;
        if first.replicate then crosspoint =int(&count_vars.*ranuni(35346)) + 1;
        *create the piece to swap;
        swap=substr(chromosone,crosspoint);
        *make counter by replicate group so as to reverse order;
        if first.replicate then option=1;
        else option = 2;
        output swappiece ;
        output crossoverg2;
run;

*sort the pieces in descending order to enable swap;
proc sort data=swappiece out=swappiece (rename=(swap=swap2));
        by replicate descending option ;
run;

*merge back on and do the swap;
*this step also does the mutation and then outputs the new;
*chromosone and the replicate number;

data crossoverg2;
        merge crossoverg2 swappiece;
        *keep a copy of old chromosone for checking;
        oldchro=chromosone;
        *make the new chromosone;
        substr(chromosone,crosspoint)=swap2;
```

```
                *now mutate one alele in .1% of cases;
                *first generate a random number between 1 and 10;
                mutrand = int(1000*ranuni(35346)) + 1;
                *arbitrarily choose 5;
                if mutrand = 5 and crosspoint < &count_vars. then do;
                        *swap the value using the previously randomly ;
                        *generated crosspoint;
                        if substr(chromosone,crosspoint,1) = '0' then
                                substr(chromosone,crosspoint,1) = '1' ;
                        else substr(chromosone,crosspoint,1)='0' ;
                end;
        run;
        *****crossover process ends here;
        *clear to start again;
        %cleanup(win_results);

        *now take the chromosone and turn it back into alleles;
        data testfit (keep = chromosone replicate option allele1-allele&allele.);
                array allele {&allele.} 3 ;
                set crossoverg2;
                do i = 1 to &allele.;
                        allele[i]=substr(chromosone,i,1);
                end;
        run;
%mend;

*call the crossover mutation macro;
%crossmut();


*terminating function if generations = 500 or avg fitness doesn't
*change for 20 generations;
%macro terminate();
*initialise macro variables;
%let fitvalue = 0;
%let laps = 0;
%let end = 0;

*keep going until there have been 20 stable generations or until there;
*have been 500 generations;
%do %until ((%eval(&laps) = &maxit) or (%eval(&end) = &stable ));
        *store the old fitness average;
        %let fitvalueold = &fitvalue;

        *need to run for half the trials after the first run;
        %let loops = %eval(&crossreps);
        *run the fitness macro;
        %fitness;

        *do the crossover and mutation;
        %crossmut;

        proc sql noprint;
        select avg(fitness) into :fitvalue
```

```
        from nextgen;
        quit;

        *what was the average fitness of the lap;
        %if (%sysevalf(&fitvalue) = %sysevalf(&fitvalueold))
                %then %let end=%eval(&end+1);
        %else %let end = 0; /*stable counter resets*/
%put ******** stable gens &end;
        %let laps=%eval(&laps+1);
%put ******** up to lap &laps;
%put ******** average fitnessin nextgen is &fitvalue ;

proc sql noprint;
        select matched_vars into :matched_vars from dset1.winning_detail_&inv_year.;
        select winning_vars into :winning_vars from dset1.winning_detail_&inv_year.;
        select count_current into :count_current from dset1.winning_detail_&inv_year.;
        select count_historical into :count_historical from
dset1.winning_detail_&inv_year.;
        select ks_agree95 into :ks_agree95 from dset1.winning_detail_&inv_year.;
        select canonlist into :canonlist from dset1.winning_detail_&inv_year.;
quit;

%put ******** variables matched total is &matched_vars;
%put ******** variable list is &winning_vars ;
%put ******** current variable count is &count_current;
%put ******** historical variable count is &count_historical;
%put ******** canonical variable agreement (KS Stat) is &ks_agree95;
%put ******** list for canonical variables is &canonlist;

%end;

*write termination point into log;
%put laps= &laps , if &maxit then the maximum iterations were reached;
%put stable = &end;

%mend;

%terminate;

*print to log again;
options notes source source2 errors=20;

%let endtime  = %sysfunc(time(),time8.);

*print the run time;
%put run was from &startime to &endtime;
```

15 Appendix 4

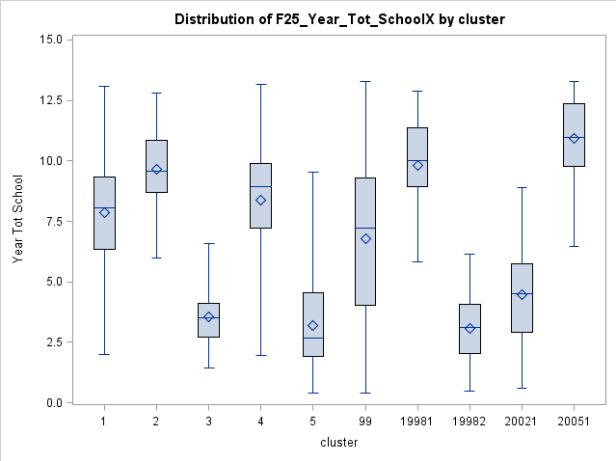
This code automatically analyses the results from a cluster analysis, including the values of the cubic clustering criterion, Pseudo F and pseudo t^2 statistics.

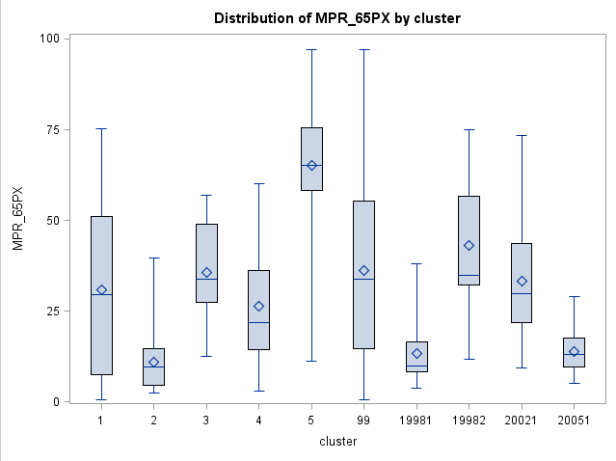
```
*keep the variables that will be used as fitness functions;
proc sort data=tree&i out=cccetc&i. (keep = _NCL_ _PSF_ _PST2_ _CCC_);
by _NCL_;
run;

*find the humps in pseudo F and the jumps in pseudo t;
*restrict to 8 clusters or less;
data cccetc (drop = first second);
set cccetc&i ;
first = lag1(_psf_);
second = lag2(_psf_);
if first > _psf_ and second < first then
PSFanal = first - second ;
else PSFanal = 0;
PST2anal=_PST2_ - lag(_PST2_);
if _ncl_ > 1 and _ncl_ <=8;

run;

*find the biggest 'hump' and 'jumps' and 'ccc';
proc rank data=cccetc descending out=cccetc;
var PST2anal PSFanal _ccc_;
run;

*select the top three of each into a macro variable;
%do j = 1 %to 3;
%let psf&j = 0;
%let pst&j = 0;
%let ccc&j = 0;

proc sql noprint;
select _ncl_ - 1 into :psf&j from cccetc where psfanal = &j;
select _ncl_ - 1 into :pst&j from cccetc where pst2anal = &j;
select _ncl_ into :ccc&j from cccetc where _ccc_= &j;
quit;
%end;

*check there is a result;
%if &psf1 = 0 and &pst1 = 0 and &ccc1 = 0
%then %do;
%let fit = 0;
%let agree = 0;
%let rsq_fit = 0; /*this is the r-squared bit below*/
%let kdvalue = 0;
%goto exit;
%end;
%else %do;
```

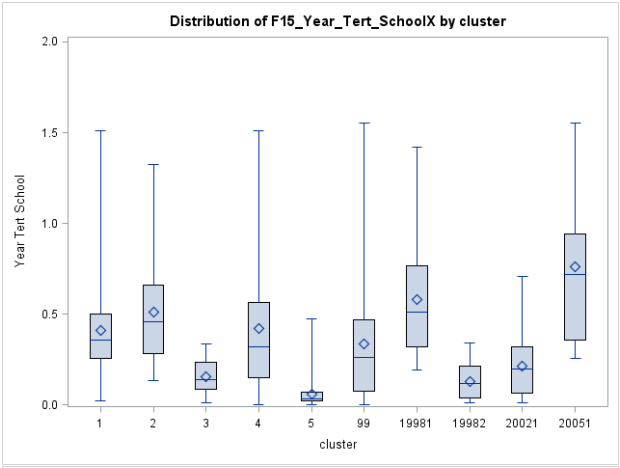
```
data score;
psf1 = &psf1; pst1 = &pst1; ccc1 = &ccc1;
psf2 = &psf2; pst2 = &pst2; ccc2 = &ccc2;
psf3 = &psf3; pst3 = &pst3; ccc3 = &ccc3;

run;
%end;

*transpose to accumulate counts. trying to ascertain;
*how much agreement there is;
proc transpose data=score out=score ; run;

*accumulate, but not if 1 cluster is selected;
proc freq data=score noprint; where col1 ne 1 and col1 ne 0; table col1 / out=score; run;

*only keep non unique values i.e. where there is agreement;
data score;
set score;
if count > 1 ;
run;

*only carry on if there are records;
*check to see if there are records;
proc sql noprint;
select count(*) into :obs_count from score;
quit;

*set fitness functions to zero and exit if there are no records;
%if &obs_count = 0 %then %do;
%let fit = 0;
%let agree = 0;
%let rsq_fit = 0; /*this is the r-squared bit below*/
%let kdvalue = 0;
%goto exit;
%end;

*store the top value;
proc sql noprint;
select max(col1) into :fit from score having count = max(count);
select max(count) into :agree from score having count = max(count);
quit;
```

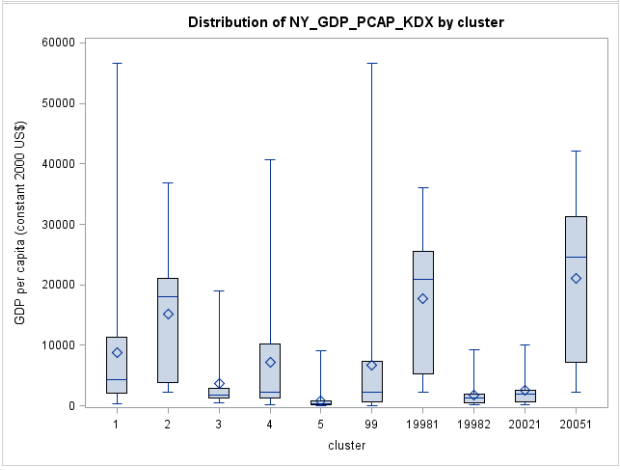
16 Appendix 5

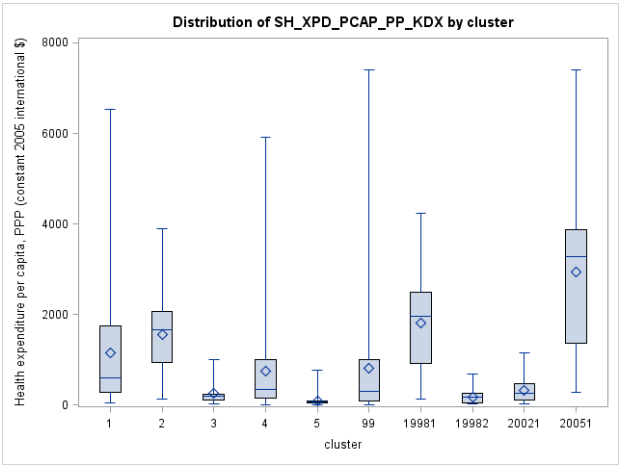
Plots used in cluster profiling. Boxplots labelled '99' refer to the whole dataset. Variable descriptions and descriptions for categorical variables can be found in Appendix 2.

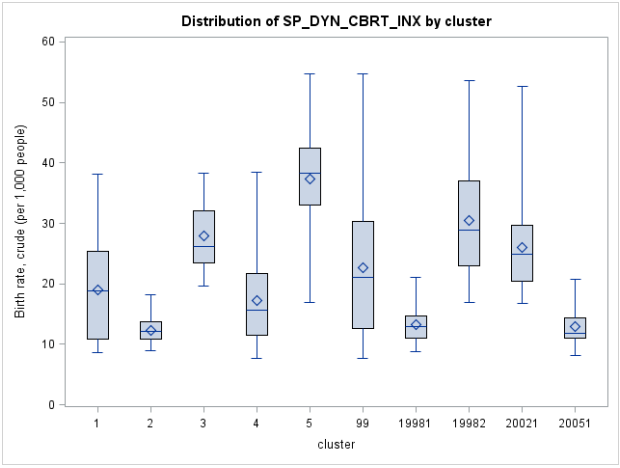


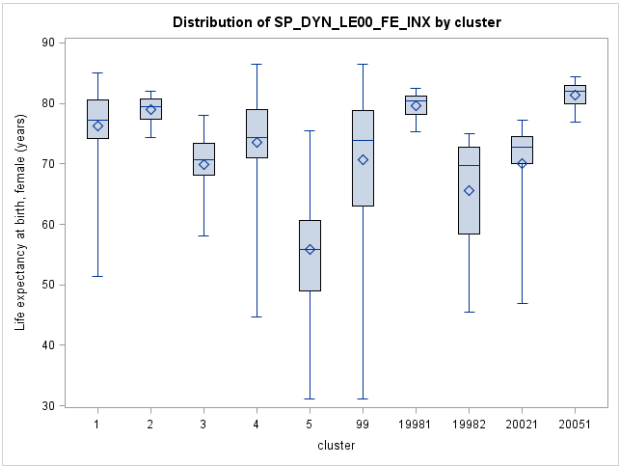


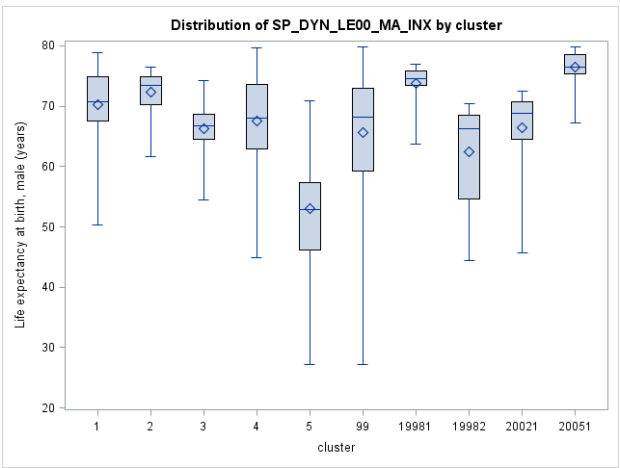

Distribution of F15_Year_Tert_SchoolX by cluster
Year Tert School
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
99
19981
19982
20021
20051
cluster

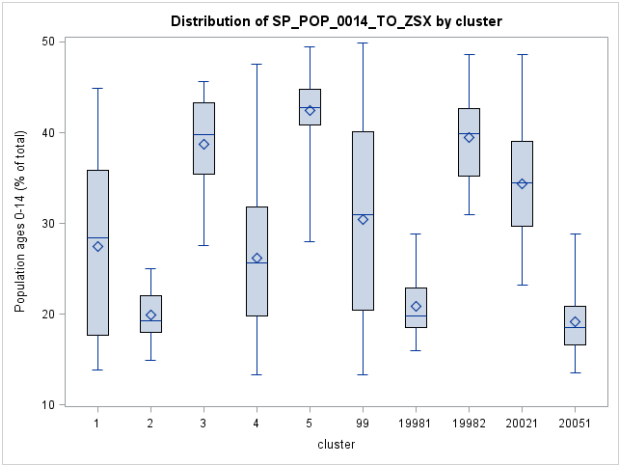

Distribution of NY_GDP_PCAP_KDX by cluster
GDP per capita (constant 2000 US$)
60000
50000
40000
30000
20000
10000
0
1
2
3
4
5
99
19981
19982
20021
20051
cluster

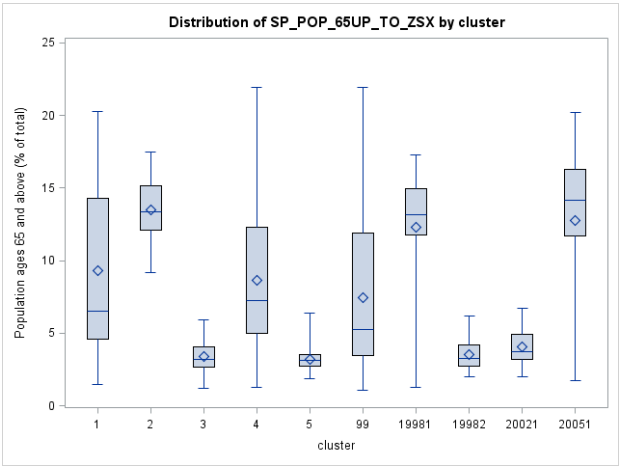

Distribution of SH_XPD_PCAP_PP_KDX by cluster
Health expenditure per capita, PPP (constant 2005 international $)
8000
6000
4000
2000
0
1
2
3
4
5
99
19981
19982
20021
20051
cluster

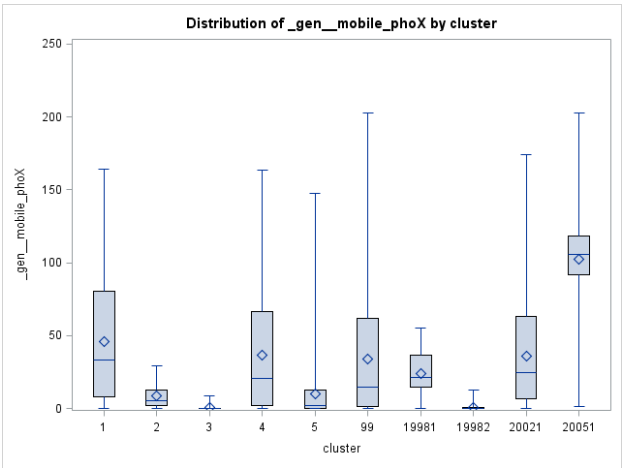

Distribution of SP_DYN_CBRT_INX by cluster
Birth rate, crude (per 1,000 people)
0
10
20
30
40
50
60
1
2
3
4
5
99
19981
19982
20021
20051
cluster

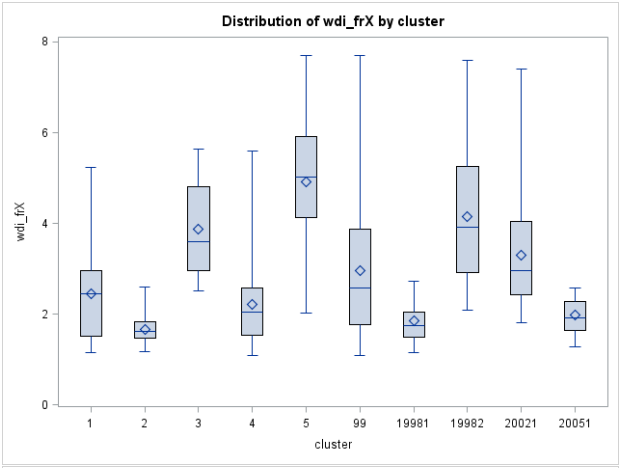

Distribution of SP_DYN_LE00_FE_INX by cluster
Life expectancy at birth, female (years)
30
40
50
60
70
80
90
1
2
3
4
5
99
19981
19982
20021
20051
cluster

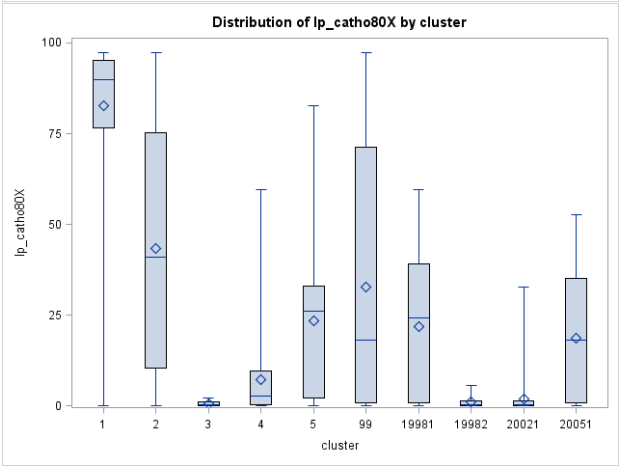

Distribution of SP_DYN_LE00_MA_INX by cluster
Life expectancy at birth, male (years)
20
30
40
50
60
70
80
1
2
3
4
5
99
19981
19982
20021
20051
cluster

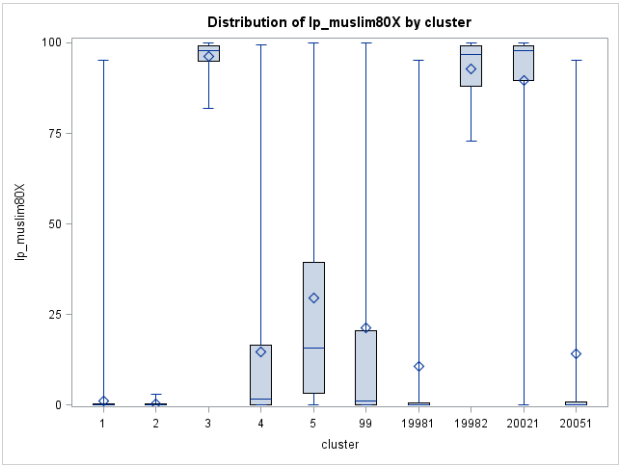

Distribution of SP_POP_0014_TO_ZSX by cluster
Population ages 0-14 (% of total)
50
40
30
20
10
1
2
3
4
5
99
19981
19982
20021
20051
cluster


Distribution of SP_POP_65UP_TO_ZSX by cluster
Population ages 65 and above (% of total)
25
20
15
10
5
0
1
2
3
4
5
99
19981
19982
20021
20051
cluster


Distribution of _gen__mobile_phoX by cluster
_gen__mobile_phoX
250
200
150
100
50
0
1
2
3
4
5
99
19981
19982
20021
20051
cluster


Distribution of wdi_frX by cluster
wdi_frX
0
2
4
6
8
1
2
3
4
5
99
19981
19982
20021
20051
cluster


Distribution of lp_catho80X by cluster
lp_catho80X
0
25
50
75
100
1
2
3
4
5
99
19981
19982
20021
20051
cluster


Distribution of lp_muslim80X by cluster
lp_muslim80X
0
25
50
75
100
1
2
3
4
5
99
19981
19982
20021
20051
cluster

Civil Liberties by Cluster

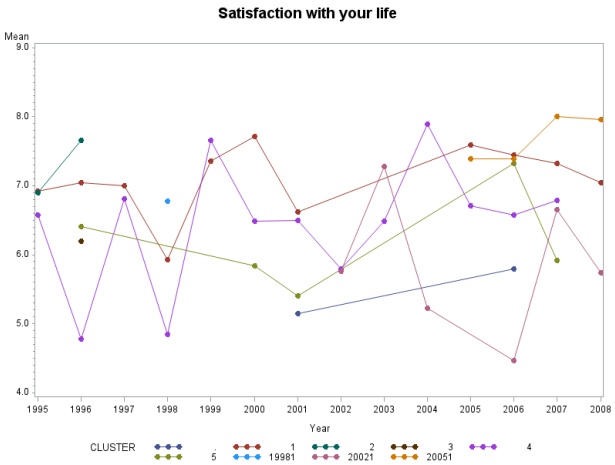
Frequency	Table of CLUSTER by fh_cl							
	fh_cl							
CLUSTER	1	2	3	4	5	6	7	Total
1	184	139	146	74	20	8	1	572
2	28	16	1	1	0	0	2	48
3	0	0	0	2	10	10	10	32
4	95	160	106	98	101	43	33	636
5	0	34	65	105	89	49	17	359
19981	11	4	0	0	1	1	1	18
19982	0	0	2	7	16	3	10	38
20021	0	4	16	18	42	23	20	123
20051	71	0	0	4	11	2	3	91
Total	389	357	336	309	290	139	97	1917
Frequency Missing = 108								

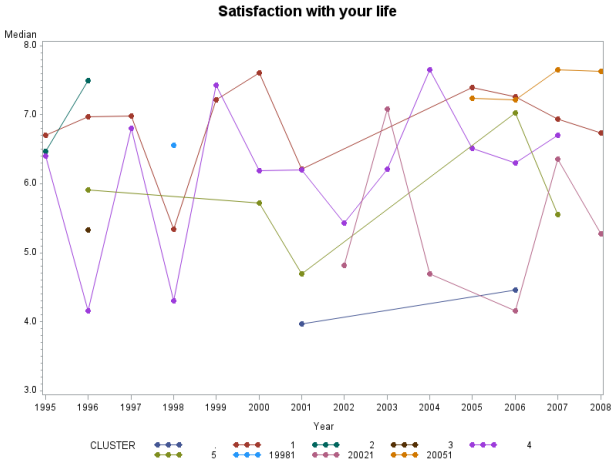
Political Rights by cluster

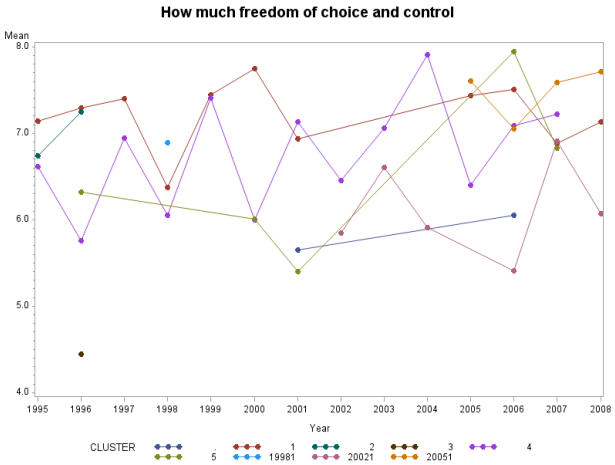
Frequency	Table of CLUSTER by fh_pr							
	fh_pr							
CLUSTER	1	2	3	4	5	6	7	Total
1	285	110	89	51	17	17	3	572
2	42	3	0	1	0	0	2	48
3	0	0	1	4	5	13	9	32
4	199	119	58	56	52	63	89	636
5	5	61	63	54	47	78	51	359
19981	15	0	0	0	1	0	2	18
19982	0	1	3	7	3	12	12	38
20021	0	11	9	5	18	46	34	123
20051	70	1	0	5	4	6	5	91
Total	616	306	223	183	147	235	207	1917
Frequency Missing = 108								

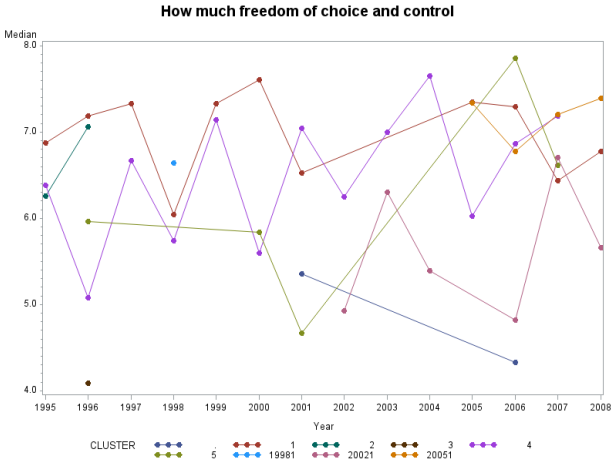
Colonisation by Cluster

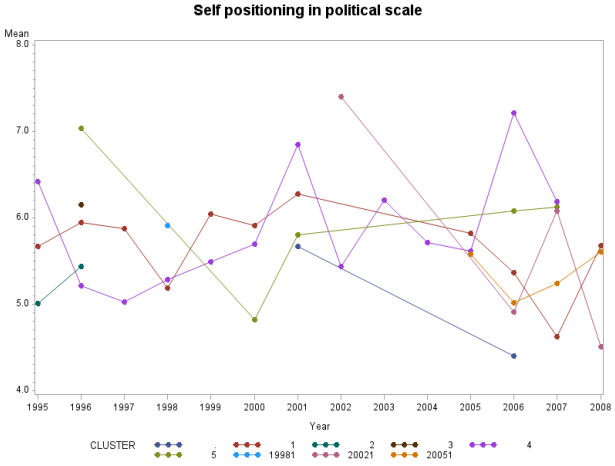
Frequency	Table of CLUSTER by ht_colonial										
	ht_colonial										
CLUSTER	-10	-8	-7	-6	-5	-4	-3	-2	-1	0	Total
1	0	0	15	36	18	15	0	249	0	239	572
2	0	0	0	0	0	0	0	4	0	44	48
3	0	0	0	10	14	0	2	0	0	6	32
4	0	0	0	34	240	0	1	10	14	339	638
5	14	30	15	122	154	0	2	0	1	21	359
19981	0	0	0	0	2	0	0	2	0	14	18
19982	0	0	0	16	14	0	2	0	0	6	38
20021	1	0	0	45	45	0	8	0	0	24	123
20051	0	0	0	0	15	0	0	5	0	71	91
Total	15	30	30	263	502	15	15	270	15	764	1919
Frequency Missing = 106											

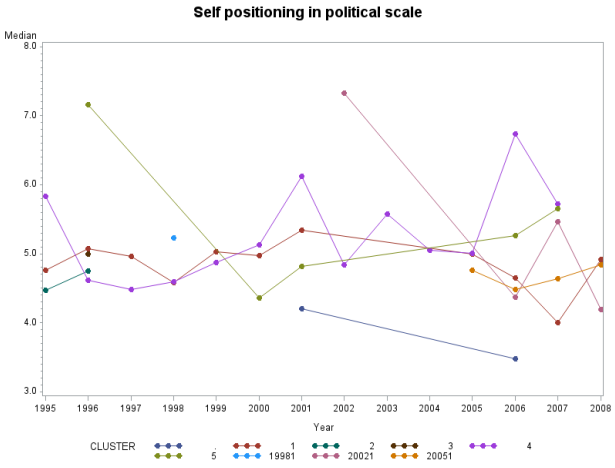

Satisfaction with your life
Mean
9.0
8.0
7.0
6.0
5.0
4.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Year
CLUSTER . 1 2 3 4 5 19981 20021 20051


Satisfaction with your life
Median
8.0
7.0
6.0
5.0
4.0
3.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Year
CLUSTER . 1 2 3 4 5 19981 20021 20051

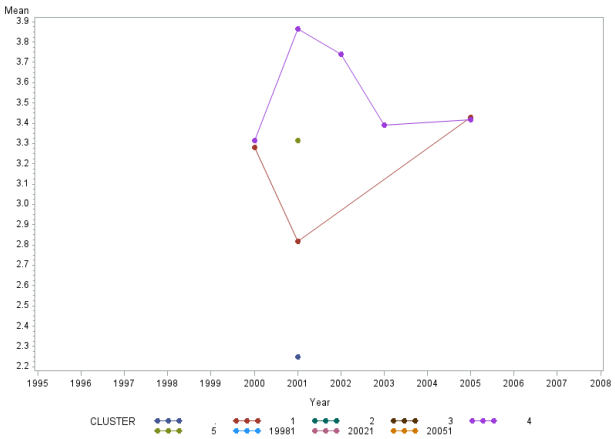

How much freedom of choice and control
Mean
8.0
7.0
6.0
5.0
4.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Year
CLUSTER . 1 2 3 4 5 19981 20021 20051


How much freedom of choice and control
Median
8.0
7.0
6.0
5.0
4.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Year
CLUSTER
1
2
3
4
5
19981
20021
20051

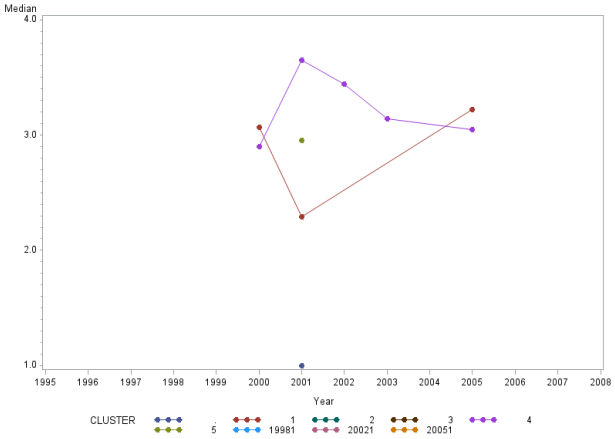

Self positioning in political scale
Mean
8.0
7.0
6.0
5.0
4.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Year
CLUSTER
1
2
3
4
5
19981
20021
20051


Self positioning in political scale
Median
8.0
7.0
6.0
5.0
4.0
3.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Year
CLUSTER
1
2
3
4
5
19981
20021
20051

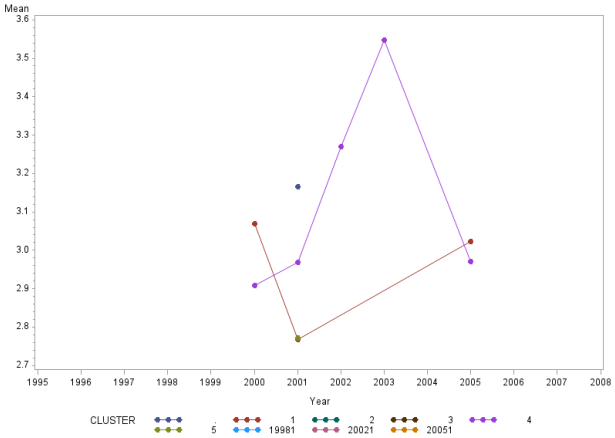
Current society: Egalitarian vs. competitive society

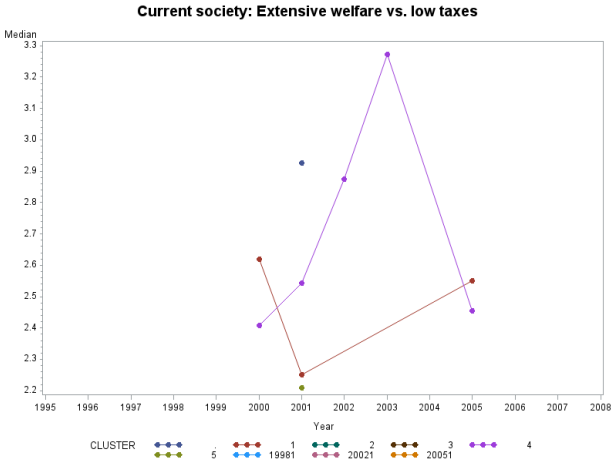


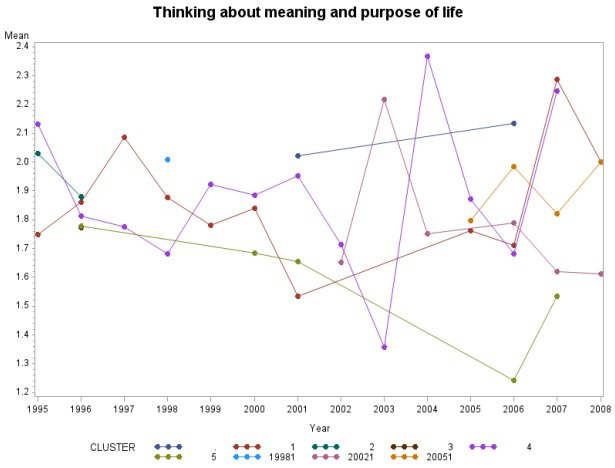
Current society: Egalitarian vs. competitive society

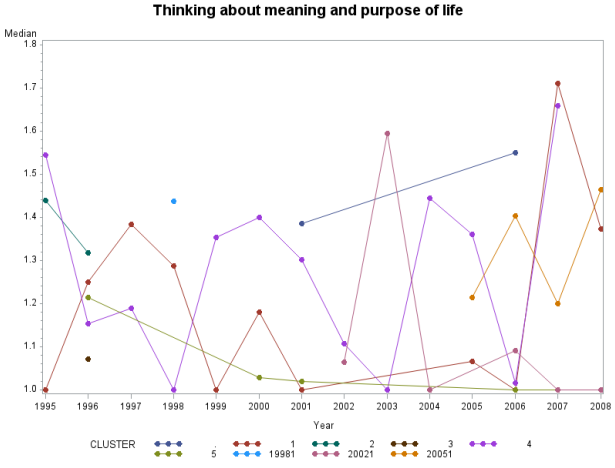


Current society: Extensive welfare vs. low taxes

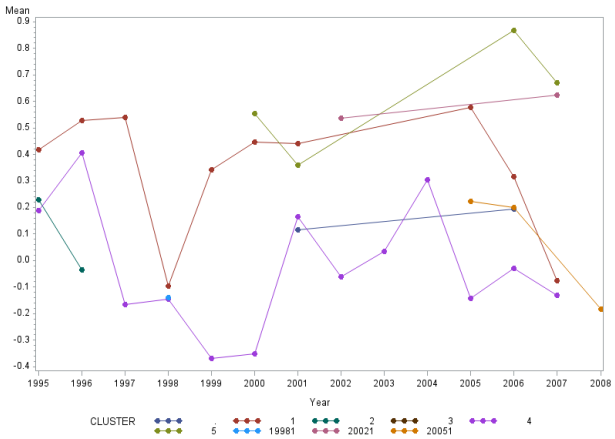



Current society: Extensive welfare vs. low taxes
Median
Year
CLUSTER
1
2
3
4
5
19981
20021
20051

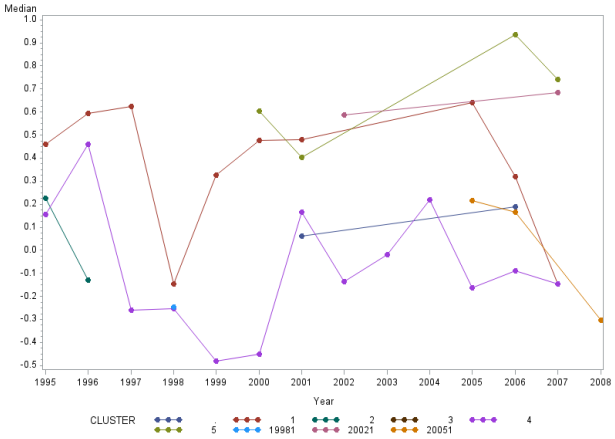

Thinking about meaning and purpose of life
Mean
Year
CLUSTER
1
2
3
4
5
19981
20021
20051


Thinking about meaning and purpose of life
Median
Year
CLUSTER
1
2
3
4
5
19981
20021
20051

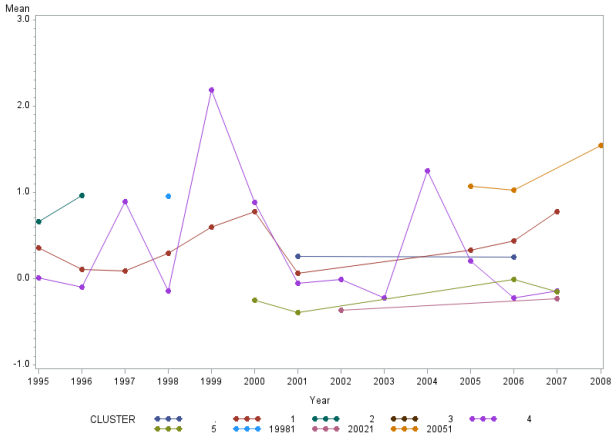
SURVIVAL/SELF-EXPRESSION VALUES



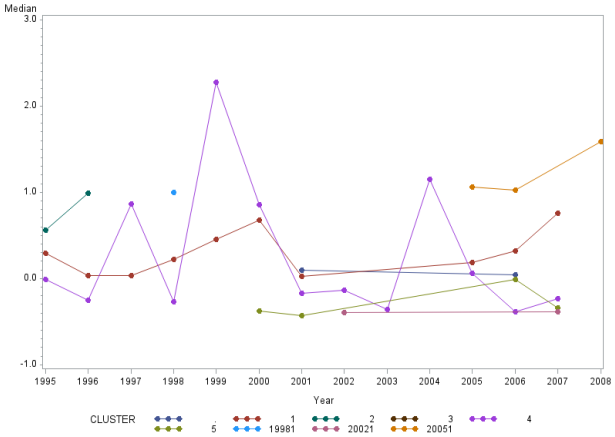
SURVIVAL/SELF-EXPRESSION VALUES



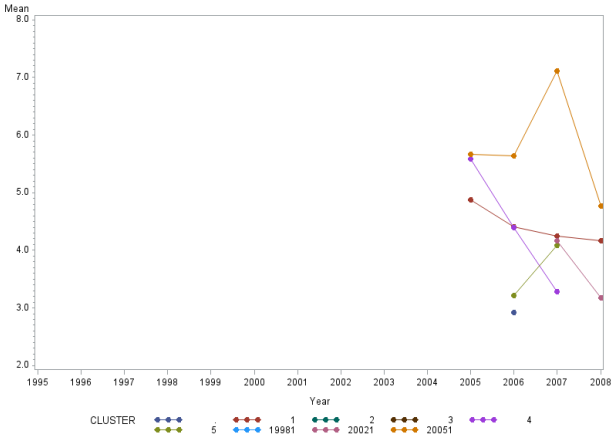
TRADITIONAL/SECULAR RATIONAL VALUES



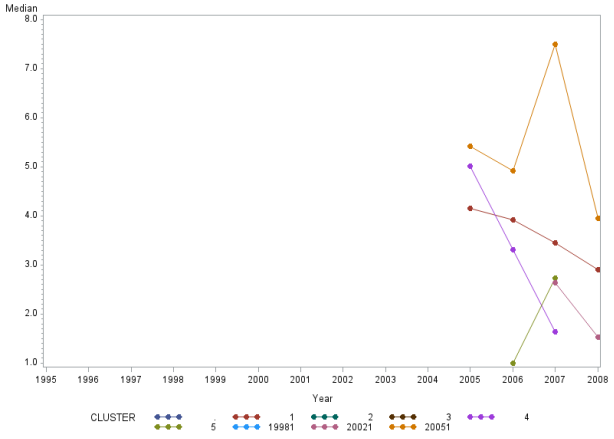
TRADITIONAL/SECULAR RATIONAL VALUES

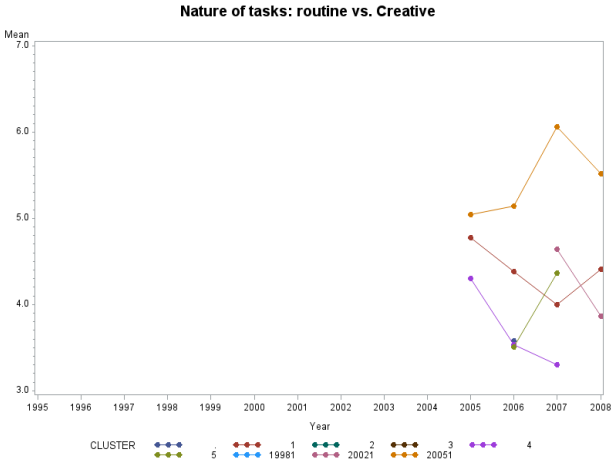


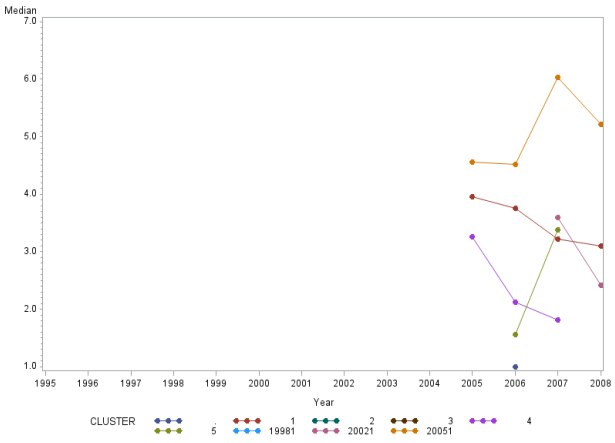
Nature of tasks: manual vs. Cognitive

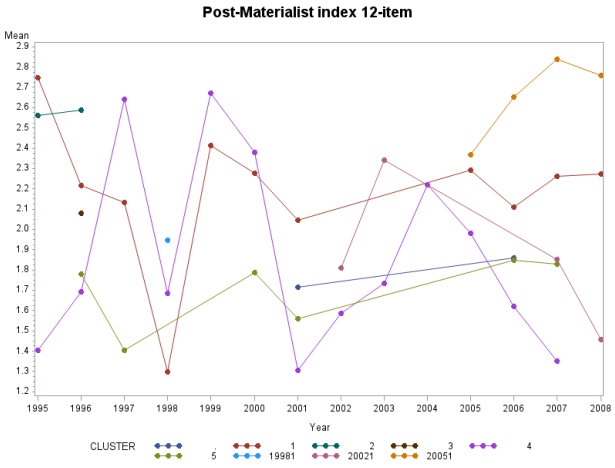


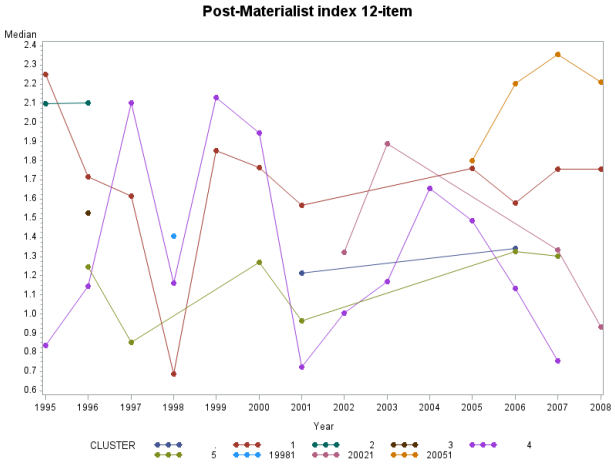
Nature of tasks: manual vs. Cognitive




Nature of tasks: routine vs. Creative
Mean
7.0
6.0
5.0
4.0
3.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Year
CLUSTER
1
2
3
4
5
19981
20021
20051


Nature of tasks: routine vs. Creative
Median
7.0
6.0
5.0
4.0
3.0
2.0
1.0
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Year
CLUSTER
1
2
3
4
5
19981
20021
20051


Post-Materialist index 12-item
Mean
2.9
2.8
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Year
CLUSTER
1
2
3
4
5
19981
20021
20051


Post-Materialist index 12-item
Median
2.4
2.3
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
Year
CLUSTER
1
2
3
4
5
19981
20021
20051

17 Appendix 6

List of countries by cluster membership and year

Country Name	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Albania	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Algeria	3	3	4	3	3	5	5	3	3	3	3	3	3	3	3
Argentina	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Armenia	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Australia	2	2	4	2	4	4	4	1	4	4	2	2	2	2	2
Austria	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Bahrain	3	3	4	2	4	4	4	1	4	4	2	2	2	2	2
Bangladesh	5	5	5	3	3	5	5	3	3	3	3	5	5	5	5
Barbados	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Belgium	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Belize	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Benin	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Bolivia	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Botswana	5	5	5	5	5	5	5	4	4	4	4	4	4	4	4
Brazil	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Brunei	4	4	4	4	4	4	4	1	4	4	2	2	2	2	2
Bulgaria	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Burundi	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Cambodia	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Cameroon	5	5	5	5	5	5	5	5	5	5	5	5	1	1	1
Canada	2	2	4	2	1	1	1	1	1	4	2	2	2	2	2
Central African Republic	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Chile	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
China	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Colombia	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Congo, Dem. Rep.	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Congo, Rep.	1	1	5	5	1	1	5	5	1	1	1	5	1	1	1
Costa Rica	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Cote d'Ivoire	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Croatia	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Cuba	2	2	4	2	4	4	4	1	4	4	2	2	2	2	2
Cyprus	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Czech Republic	2	2	4	2	4	4	4	1	4	4	2	2	2	2	2
Denmark	2	2	4	2	4	4	4	1	4	4	2	2	2	2	2
Dominican Republic	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ecuador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Egypt	3	3	4	3	3	5	5	3	3	3	3	3	3	3	3

El Salvador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Estonia	2	2	4	2	4	4	4	1	4	4	2	2	2	2	2
Fiji	4	4	5	4	4	4	4	4	4	4	4	4	4	4	4
Finland	2	2	4	2	4	4	4	1	4	4	2	2	2	2	2
France	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Gabon	1	1	5	5	1	1	5	5	1	1	1	5	1	1	1
Gambia	5	5	4	4	4	4	4	4	4	4	4	4	4	4	4
Germany	2	2	4	2	1	1	1	1	1	4	2	2	2	2	2
Ghana	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Greece	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Guatemala	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Guyana	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Haiti	1	1	5	5	1	1	1	1	1	1	1	1	1	1	1
Honduras	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Hungary	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1
Iceland	2	2	4	2	4	4	4	1	4	4	2	2	2	2	2
India	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Indonesia	4	4	5	4	4	4	4	4	4	4	4	4	4	4	4
Iran	3	3	4	3	3	5	5	3	3	3	3	3	3	3	3
Iraq	3	3	4	3	3	5	5	3	3	3	3	3	3	3	3
Ireland	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Israel	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Italy	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Jamaica	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Japan	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Jordan	3	3	4	3	3	5	5	3	3	3	3	3	3	3	3
Kazakhstan	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Kenya	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Korea, South	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Kuwait	3	3	4	2	4	4	4	1	4	4	2	2	2	2	2
Kyrgyzstan	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Laos	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Latvia	4	4	4	4	4	4	4	4	4	4	4	4	4	4	2
Lesotho	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Liberia	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Libya	3	3	4	3	3	5	5	3	3	3	3	3	3	3	3
Lithuania	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Luxembourg	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Malawi	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Malaysia	4	4	5	4	4	4	4	4	4	4	4	4	4	4	4
Maldives	3	3	5	3	3	5	5	3	3	3	3	5	5	5	5
Mali	5	5	5	3	3	5	5	3	3	3	3	5	5	5	5
Malta	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Mauritania	3	3	5	3	3	5	5	3	3	3	3	5	5	5	5
Mauritius	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Mexico	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Moldova	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Mongolia	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Morocco	3	3	4	3	3	5	5	3	3	3	3	3	3	3	3

Mozambique	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Myanmar	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Namibia	4	4	5	4	4	4	4	4	3	3	3	5	5	5	5
Nepal	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Netherlands	2	2	4	2	1	1	1	1	1	4	2	2	2	2	2
New Zealand	2	2	4	2	4	4	4	1	4	4	2	2	2	2	2
Nicaragua	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Niger	5	5	5	3	3	5	5	3	4	4	4	4	4	4	4
Norway	2	2	4	2	4	4	4	1	4	4	2	2	2	2	2
Pakistan	3	3	5	3	3	5	5	3	3	3	3	5	5	5	5
Panama	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Papua New Guinea	5	5	5	5	5	5	5	5	3	5	5	5	5	5	5
Paraguay	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Peru	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Philippines	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Poland	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Portugal	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Qatar	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Romania	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Russia	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Rwanda	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Saudi Arabia	3	3	4	3	3	5	5	3	3	3	3	3	3	3	3
Senegal	5	5	5	3	3	5	5	3	3	3	3	5	5	5	5
Sierra Leone	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5
Singapore	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Slovakia	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Slovenia	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
South Africa	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Spain	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sri Lanka	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Sudan	5	5	5	3	3	5	5	3	3	3	3	5	5	5	5
Swaziland	4	4	5	4	4	4	4	4	4	4	4	4	4	4	4
Sweden	2	2	4	2	4	4	4	1	4	4	2	2	2	2	2
Switzerland	2	2	4	2	1	1	1	1	1	4	2	2	2	2	2
Syrian Arab Republic	3	3	4	3	3	5	5	3	3	3	3	3	3	3	3
Tajikistan	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Thailand	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Togo	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Tonga	4	4	5	4	4	4	4	4	4	3	3	5	4	4	4
Trinidad and Tobago	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Tunisia	3	3	4	3	3	5	5	3	3	3	3	3	3	3	3
Turkey	3	3	4	3	3	5	5	3	3	3	3	3	3	3	3
Uganda	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Ukraine	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
United States of America	2	2	4	2	4	4	4	1	4	4	2	2	2	2	2
Uruguay	2	2	4	2	4	4	4	1	1	1	1	1	1	1	1
Venezuela	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Vietnam	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Zambia	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

1-Traditional Forward Thinkers
2- Satisfied Free and Central
3-Religious Traditionals
4-Middling and Competitive
5-Struggling Traditionals